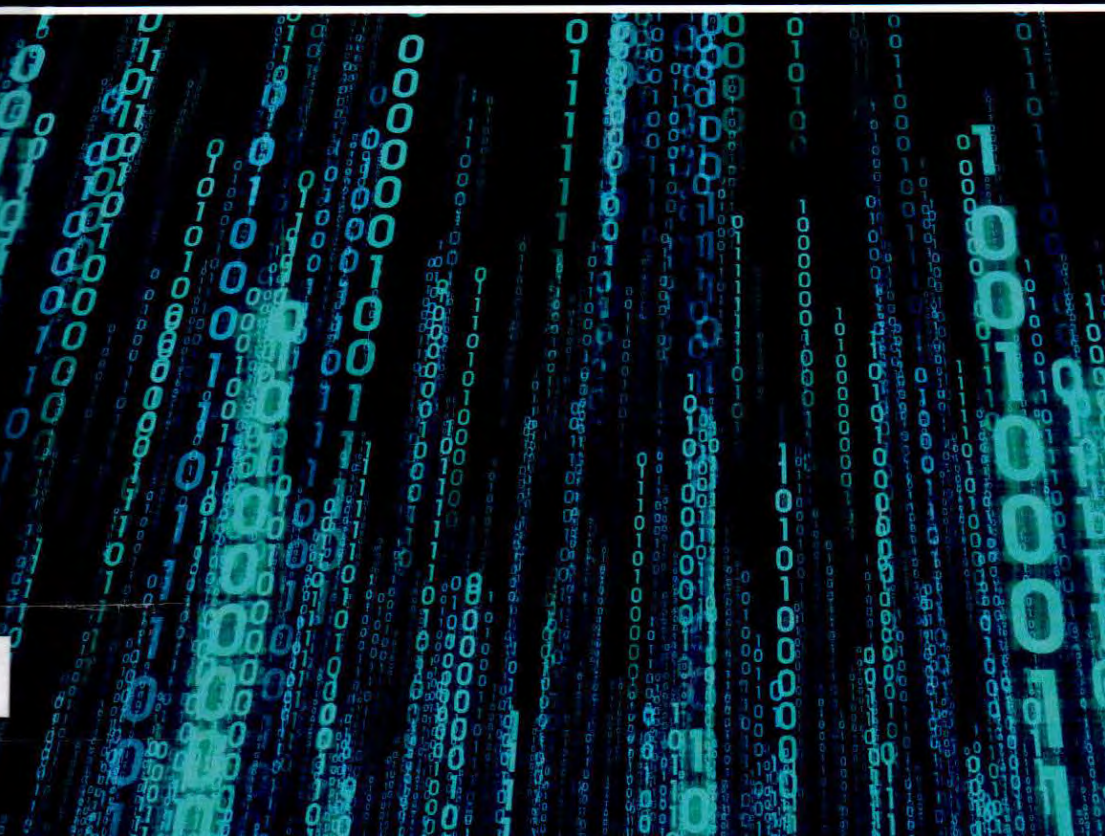
NATIONAL GEOGRAPHIC

5

LA INTELIGENCIA ARTIFICIAL

El camino hacia la ultrainteligencia



La inteligencia artificial experimenta ya un auge sin precedentes, pero el próximo paso en su desarrollo supone uno de los mayores retos tecnológicos de la historia: el diseño de máquinas con una inteligencia equiparable a la del ser humano. Organizadas como redes de neuronas artificiales, serán capaces de registrar y analizar, mediante algoritmos cada vez más complejos, los terabytes de información de los que disponemos actualmente, para tomar decisiones y actuar de forma racional. Pero ¿podrán perfeccionarse a sí mismas y crear ultrainteligencias todavía más complejas? ¿Estamos ante nuestra última generación como especie dominante?

Sergio Parra es periodista y divulgador científico.
Marc Torrens es ingeniero informático y doctor en inteligencia artificial por la Escuela Politécnica Federal de Lausana.

LA INTELIGENCIA ARTIFICIAL

El camino hacia
la ultrainteligencia

© Sergio Parra y Marc Torrens por el texto
© 2017, RBA Coleccionables, S.A.U.

Realización: EDITEC

Diseño cubierta: Llorenç Martí

Diseño interior: tactilestudio

© Ilustraciones: Francisco Javier Guarga Aragón

Fotografías: Shutterstock: cubierta; Age Fotostock: 45, 57, 101bi; John McCarthy's home page: 49ai; Frank Rosenblatt: 49ad; Carnegie Mellon University: 49b; The RobotCub project: 65; Getty Images: 101a; Archivo RBA: 83; Daderot: 101bd.

ISBN (OC): 978-84-473-9071-7
ISBN: 978-84-473-9081-6
Depósito legal: B.23658-2017

Impreso en Liberdúplex

Impreso en España - *Printed in Spain*

Para México

Edita

RBA Editores México, S. de R.L. de C.V. Av. Patriotismo 229, piso 8, Col. San Pedro de los Pinos, CP 03800, Deleg. Benito Juárez, Ciudad de México, México

Fecha primera publicación en México: mayo 2018.

Editada, publicada e importada por RBA Editores México, S. de R.L. de C.V. Av. Patriotismo 229, piso 8, Col. San Pedro de los Pinos, CP 03800, Deleg. Benito Juárez, Ciudad de México, México

Impresa en Liberdúplex, Crta. BV-2249, Km. 7.4, Pol. Ind. Torrentfondo 08791 Sant Llorenç d'Hortons, Barcelona

ISBN: 978-607-9495-24-4 (Obra completa)
ISBN: 978-607-9495-43-5 (Libro)

Reservados todos los derechos. Ninguna parte de esta publicación puede ser reproducida, almacenada o transmitida por ningún medio sin permiso del editor.

SUMARIO

Introducción 7

01 **El amanecer de la era de las máquinas** 13

02 **De Turing al *big data*** 39

03 **Máquinas que razonan e interactúan** 69

04 **Máquinas que aprenden** 103

05 **El futuro es hoy. La IA en el mundo real** 123

Lecturas recomendadas 139

Índice 141

INTRODUCCIÓN

Cuando nació la imprenta gracias al alemán Johannes Gutenberg hacia 1439, los cambios sociales que se produjeron fueron tan espectaculares como los que hoy en día está propiciando internet. A grandes rasgos, una máquina, la imprenta, amplificó las capacidades cognitivas del cerebro humano. Así, hoy los cerebros lectores entienden de otra forma el lenguaje, procesan de manera diferente las señales visuales, e incluso razonan y forman los recuerdos de otro modo. Leer nos permitió penetrar en la mente de personas que nunca hubiéramos conocido. A pesar de esta revolución que supuso, o tal vez a causa de ella, fueron muchos los que abominaron de la imprenta. Algunos la acusaron de poner en manos del vulgo conocimientos que no le convenían, mientras que otros, como el naturalista y bibliógrafo suizo Conrad Gessner, expresaron su inquietud por la sobrecarga de información que supondría, adelantándose cinco siglos a las acusaciones que se vertieron sobre internet en la misma línea. Pero es que las innovaciones tecnológicas raramente evolucionan como anticipan sus contemporáneos: la radio no mató la conversación, el teléfono no esquilmó la intimi-

dad, ni los ordenadores personales interesaron tan solo a un simple puñado de aficionados a la informática.

Muchas de las interpretaciones más pesimistas sobre los efectos de una nueva invención nacen del recelo que despierta todo cambio vertiginoso, sobre todo cuando es de índole tecnológica. Lo cierto es que las variables y ramificaciones de una revolución tecnológica suelen ser extraordinariamente complejas. Además, la tecnología no es algo que aparezca sin más en un contexto dado; más bien se integra con él y lo redefine. El historiador de la tecnología Chris Otter lo resume así: «los valores victorianos como la puntualidad, la pulcritud y la atención fueron efectos colaterales de la creación de los relojes precisos, el agua corriente y las gafas».

Esta reflexión viene a cuento porque la tecnología objeto de este libro, la inteligencia artificial, o IA, posee un potencial transformador tal vez único en la historia. Y, por consiguiente, despierta recelos proporcionales a dicho potencial. Desde los asistentes personales activados por voz, que se crean un espacio cada vez mayor en nuestra vida diaria, a las redes neuronales que navegan sin descanso por el *big data* (macrodatos) en busca de patrones que algún día sirvan para predecir nuestros más ínfimos deseos, los programas de IA están ya transformando grandes áreas de la economía y la vida cotidiana. Sin embargo, con todo su poder, no dejan de ser aplicaciones de una inteligencia limitada a tareas concretas en ámbitos concretos. Pero ¿qué ocurriría si alguna vez se creara una IA con capacidades plenamente equiparables a las humanas? Estas IA serían capaces de diseñar máquinas más inteligentes que ellas mismas, en una suerte de escalada cognitiva que pondría al *Homo sapiens* ante el reto de compartir el planeta con otra especie dominante. Según un número considerable de científicos, tecnólogos y futuristas, la emergencia de estas ultrainteligencias artificiales conduciría a un punto de no retorno tecnológico conocido como *singularidad*. Aunque resulta difícil anticipar las consecuencias de un evento de este calibre, sus potenciales beneficios a escala planetaria serían inmensos.

Ante el vértigo que provoca tal visión desde las alturas es posible que se nos olvide que antes hay que conseguir emular la inteligencia humana. El principal escollo al que se enfrentan los investigadores es que todavía no se dispone de un modelo completo de la cognición humana ni en lo tocante a los procesos conscientes ni en lo relativo a sus correlatos neurológicos. Sin embargo, los adelantos en el estudio de la cognición, la neurociencia y la neurociencia computacional, unidos al continuo progreso técnico en computación, hacen que los expertos en la materia sean optimistas en alcanzar una IA de nivel humano este mismo siglo. Ahora bien, la propia idea de una IA de nivel humano acarrea problemas no solo tecnológicos, sino también filosóficos. ¿Cómo estar seguros de que una máquina es consciente y actúa según su libre albedrío? Podría ser que lo que tomamos por inteligencia no fuera más que un programa de suma complejidad ante cuyas exigencias la IA tiene tan poca libertad como el menú de un cajero automático. El test de Turing, la más conocida de las pruebas para determinar si una inteligencia artificial es equiparable a la humana, sencillamente ignora el problema. Para Alan Turing, su creador, no hay diferencia importante entre un ser inteligente y otro que lo imite a la perfección. Nunca seremos capaces de meternos en la cabeza de nadie ni de nada, ya sea un ser vivo o una máquina, para saber si es consciente o si actúa según su libre albedrío. La mayoría de los científicos aceptan tácitamente esta premisa, y se concentran en resolver las cuestiones tecnológicas.

Fue precisamente una serie de artículos de Turing a mediados de los años cincuenta del siglo pasado la que inauguró el campo de estudio de la IA. Poco después, en 1956, se celebró en la Universidad de Dartmouth un encuentro donde se definieron dos líneas de investigación que dominaron la disciplina en sus primeras décadas: la simbólica y la conexionista. La primera, liderada por figuras como Marvin Minsky, Herbert A. Simon o Allen Newell, postula que para reproducir la actividad cognitiva del cerebro debe partirse de modelos abstractos expresados en símbolos. Los primeros

frutos de este enfoque fueron programas capaces de reproducir actividades inteligentes muy formalizadas como, por ejemplo, obtener demostraciones matemáticas. La segunda escuela, en la que destacó Frank Rosenblatt, parte de la base de que la inteligencia es una propiedad funcional del cerebro biológico y que para simularla el mejor camino es reproducir la estructura de aquel. Ya en su arranque, el modelo conexionista dio a luz conceptos que se han demostrado cruciales, como el de red neuronal artificial, o RNA. Las RNA constituyen los primeros programas de la historia capaces de aprender de forma autónoma.

Sin embargo, tras estos espectaculares avances tempranos, una y otra escuela se toparon con importantes dificultades de índole tanto teórica como tecnológica y la disciplina entró en una fase de estancamiento, conocida como el invierno de la IA, que duró toda la década de 1970 y parte de la siguiente. La mayor sofisticación en el diseño de redes neuronales y el éxito cosechado por los sistemas expertos, programas de búsqueda y razonamiento que eran capaces de ofrecer rendimientos similares al humano en entornos especializados como el del diagnóstico médico, permitieron a la disciplina atraer nuevamente fondos. Ello condujo a una serie de acontecimientos históricos que llevó a la IA a la primera página de todos los diarios: en 1997, el ordenador Deep Blue derrotaba en un torneo de ajedrez al gran Garri Kaspárov y en 2015, el programa informático AlphaGo hacía lo propio con Lee Sedol, campeón de go. Entre tanto, el sistema informático Watson se imponía en el concurso televisivo estadounidense Jeopardy! En la actualidad, la IA es un elemento habitual en las aplicaciones y servicios de la economía digital, desde Siri a Facebook, y desempeñará un papel fundamental en las revoluciones que vienen, como el Internet de las cosas, el vehículo autónomo o la irrupción de los robots en la vida cotidiana.

Aunque la disciplina ha desbordado sus cauces históricos en términos de alcance y financiación, sus áreas de interés se han mantenido históricamente estables y nos permiten trazar tres grandes

ámbitos de investigación: el razonamiento y la adquisición de conocimiento; la comunicación e interacción con el entorno, incluidos los seres vivos, y el aprendizaje. En el primer ámbito, el objetivo es simular el razonamiento y la toma de decisiones en condiciones de información imperfecta y ambigua, así como la adecuada representación y organización del conocimiento para, mediante sofisticadas técnicas de búsqueda, extraer la información relevante con la que alimentar el proceso razonador. En el segundo se persigue que la máquina pueda ver e interpretar objetos, así como entender el lenguaje natural y comunicarse con seres humanos gracias a él. Se trata de un campo de investigación en el que la robótica ha irrumpido con fuerza gracias al convencimiento de que la inteligencia puede ser inseparable de unos sentidos y un cuerpo físico con el que relacionarse con el entorno. Por último, el ámbito del aprendizaje está experimentando una auténtica revolución gracias a la creciente sofisticación de redes neuronales cada vez más potentes. Las RNA actuales son capaces de extraer sutiles pautas en los datos y aplicarlos al reconocimiento de imágenes y sonidos o a la predicción de nuestro comportamiento en los ámbitos más variados. A esta revolución ha contribuido el enorme rastro digital que dejamos como usuarios cada vez más monitorizados de internet o las redes sociales. A medida que estas redes neuronales pueden ser alimentadas de datos cada vez más variados se aproximarán al modelo de un niño que aprende. En este punto se habrá satisfecho una aspiración que se remonta al propio Turing, cuando sugirió que, en lugar de partir de una mente adulta, el camino más rápido hacia la IA podría ser modelar una mente infantil y educarla.

Armados de todo este conocimiento, nos hallamos en una mejor posición a la hora de reconocer y valorar importantes patrones en la evolución tecnológica de la sociedad. Porque conceptos tales como ultrainteligencia o singularidad alimentan la imaginación, pero a la espera de que acontezcan, si es que algún día lo hacen, corremos el riesgo de no apreciar un evento no menos histórico, apasionante y transformador: el despliegue de la IA en nuestras vidas.

01

EL AMANECER DE LA ERA DE LAS MÁQUINAS

Si fuéramos capaces de crear una inteligencia artificial equiparable a la humana, se desencadenaría un cambio radical para la humanidad. Al otro lado de este punto de inflexión, conocido como *singularidad*, nos esperaría un futuro dominado por máquinas ultrainteligentes. ¿Qué hay de plausible en esta hipótesis? ¿Cuál sería nuestro rol como especie en esta nueva era?

Tradicionalmente se ha considerado que el ser humano se encuentra en la cúspide de la evolución biológica, al menos en términos de inteligencia. Sin embargo, según algunos científicos y tecnólogos se acerca el momento en que podremos vernos superados por una inteligencia artificial fruto de nuestra propia tecnología. Esta inteligencia artificial, o IA, sería capaz de mejorarse a sí misma de forma creciente. El resultado de esta explosión de inteligencia sería una inteligencia no humana de una capacidad insospechada. El matemático británico Irving John Good, colega de Alan Turing en el laboratorio militar de Bletchley Park durante los años que vieron nacer los primeros ordenadores, escribió en 1965 un famoso párrafo al respecto:

> Definamos una máquina ultrainteligente como aquella que puede superar la capacidad intelectual de todo ser humano en no importa qué actividad. Como el diseño de máquinas es una de estas actividades, una máquina ultrainteligente sería capaz de diseñar máquinas todavía mejores; habría una explosión de inteligencia y el intelecto

> del hombre quedaría muy atrás. En consecuencia, la primera máquina ultrainteligente será el último invento que el hombre deba descubrir (...)

Los adelantos tecnológicos al alcance de esta ultrainteligencia serían de tal calibre que sus consecuencias humanas y sociales escapan a toda estimación presente. Varios científicos y pensadores, desde el matemático John von Neumann en la década de 1950 hasta su colega Vernor Vinge más recientemente, escogieron el término *singularidad* para referirse a este punto de inflexión en la historia humana más allá del cual el futuro se torna impredecible.

Estas ideas han encontrado eco en personalidades como el tecnólogo y empresario estadounidense Raymond Kurzweil, o su socio, el ingeniero y físico Peter Diamandis, reconocidos por sus aportaciones tecnológicas y sus predicciones futuristas.

En particular Kurzweil, Medalla Nacional de Tecnología por sus trabajos pioneros en reconocimiento de voz, ha dedicado varios libros —*La era de las máquinas espirituales* (1999), *La singularidad está cerca* (2005)— a la singularidad tecnológica y sus posibles consecuencias. Su conclusión, radicalmente positiva, es que «la singularidad nos permitirá trascender las limitaciones de nuestros cuerpos y cerebros biológicos. No habrá distinción, después de la singularidad, entre el ser humano y la máquina». En 2008, Diamandis y Kurzweil fundaron, con el apoyo de Google y la NASA, la Singularity University, una institución dedicada a formar a líderes y ejecutivos en el desarrollo de las tecnologías exponenciales, es decir, las sujetas, según ellos, a crecimientos de tipo exponencial, como la biotecnología, la nanotecnología y, naturalmente, la IA. El objetivo: orientar y guiar estas herramientas para resolver los grandes desafíos de la humanidad.

Porque, una vez alcanzada la singularidad, cabe preguntarse: ¿qué objetivos y motivaciones tendrían estas máquinas ultrainteligentes? Y, más importante aún, ¿seríamos capaces de contro-

larlas? Desde una perspectiva distinta, la de la filosofía aplicada, el sueco Nick Bostrom, director del Instituto para el Futuro de la Humanidad en la Universidad de Oxford, encuentra razones para la prudencia en su obra *Superinteligencia: caminos, peligros, estrategias* (2014), hasta el punto de afirmar que una superinteligencia artificial podría llegar a suponer un peligro para el ser humano mayor que cualquier otro invento de la historia. Bostrom escribe:

> Ante la perspectiva de una explosión de la inteligencia, nosotros los humanos somos como niños que juegan con una bomba. Tal es la desproporción entre el poder de nuestro juguete y la inmadurez de nuestra conducta. La superinteligencia es un reto para el que no estamos preparados y no lo estaremos hasta de aquí a mucho tiempo. No tenemos idea de cuándo se va a producir la explosión, pero si nos acercamos el aparato al oído podemos oír un leve tic-tac, tic-tac.

Bostrom justifica su prudencia sobre la base de dos tesis. La primera, llamada *tesis de ortogonalidad*, afirma que la inteligencia, entendida como la habilidad para «la predicción, la planificación y el razonamiento de medios-fines en general», es compatible con muchos y muy distintos fines, es decir, que no cabe suponer que un ente ultrainteligente tendrá, solo por serlo, fines compatibles con los seres humanos o incluso comprensibles por ellos. La segunda tesis, la de la *convergencia instrumental*, complementa la primera y sostiene que, no obstante la extrema variabilidad de fines posibles de una ultrainteligencia, cabe esperar que comparta una serie de valores tales como la autoconservación, la automejora y el interés por adquirir recursos, o sea, valores que aumentan la capacidad de alcanzar fines, sean cuales sean. Si ambas tesis fueran correctas, sostiene Bostrom, podría producirse lo que denomina un «giro traicionero»: la ultrainteligencia, en razón de los valores instrumentales antes definidos, orienta su poder a reunir todos los recursos disponibles, para así garantizar su conservación y mejora constante.

Una vez alcanzado este primer objetivo, se emplearía en alcanzar sus fines sin que nadie pudiera detenerla. Incluso en el caso de que dichos fines no fueran intrínsecamente perversos, habría muchas formas en las que una ultrainteligencia podría perseguir fines no deseables. Imaginemos por un momento que los programadores fijan una meta específica, como es la de hacernos sonreír, y que la ultrainteligencia descubre una manera de alcanzar ese objetivo final, por ejemplo, paralizando la musculatura facial para producir sonrisas de manera constante. La ultrainteligencia, sin duda, ha conseguido el propósito de quienes la programaron, pero a costa de nuestros propios intereses como seres humanos. Replantear la meta inicial sigue pudiendo producir efectos perversos. Así, por ejemplo, si la petición adopta la forma «hacernos felices», la ultrainteligencia puede decidir hacerlo mediante la excitación forzosa de los centros de placer de nuestro cerebro. Otro tipo de giro perverso podría resultar de una ultrainteligencia que dedicara tal cantidad de recursos a alcanzar un fin que el resultado fuera catastrófico para la raza humana, como en el caso de que se le encomendara maximizar la producción de, por ejemplo, clips, y procediera a convertir grandes partes del planeta en... clips.

> Mi visión a largo plazo es que probablemente acabemos en una situación o bien muy mala o bien muy buena, no en una intermedia.
>
> NICK BOSTROM, SOBRE LOS EFECTOS DE LA ULTRAINTELIGENCIA

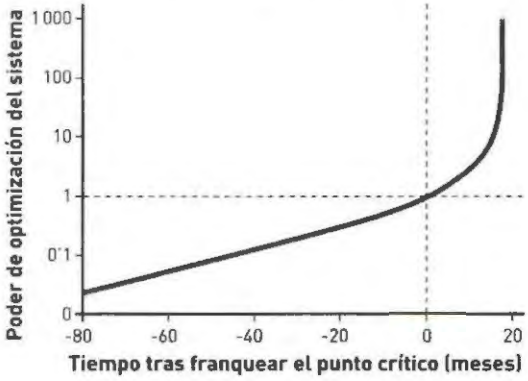
Para solucionar el problema, o al menos minimizarlo, Bostrom sugiere que el diseño de la ultrainteligencia tenga en cuenta su peligrosidad potencial e incorpore algún tipo de control. Entre esos controles los habría que afectarían a la capacidad de la ultrainteligencia, confinándola, atrofiándola de algún modo o incorporando algún tipo de sistema de apagado automático; o aquellos que buscarían incentivarla para que no desarrollara comportamientos hostiles o contraproducentes, por ejemplo, mediante órdenes di-

> UN MODELO PARA LA RAPIDEZ DE UNA EXPLOSIÓN DE INTELIGENCIA

Una vez alcanzado el objetivo de una IA humana, ¿cuán rápido pasaría esta IA a transformarse en ultrainteligencia? Para calcularlo, Bostrom propone un sencillo modelo con dos variables:

Ratio de cambio en la inteligencia = Poder de optimización / Resistencia

Donde *poder de optimización* recoge el esfuerzo efectivo dedicado a la mejora del sistema, y *resistencia* la dificultad de dicho esfuerzo en producir la mejora buscada. Hay al menos tres formas de optimizar la inteligencia de una IA. El primero serían mejoras en el *software*, como por ejemplo mejores algoritmos. El segundo serían mejoras en el *hardware*, tales como procesadores más rápidos. El tercero serían mejoras en el contenido, por ejemplo, en la calidad de los datos. Para estimar el poder de optimización, Bostrom toma la velocidad con la que crece la capacidad de cómputo de los ordenadores, la cual se dobla aproximadamente cada 18 meses. En el momento en el que la IA cruza el umbral de la inteligencia humana, este ratio se acelera porque ahora puede aplicar su intelecto a su propia mejora. El resultado es que la inteligencia del sistema se dobla cada 7,5 meses. A los 17,9 meses, se ha multiplicado por mil y podemos concluir que ha alcanzado la ultrainteligencia.



rectas o indirectas o controlando su crecimiento de forma que no evolucionara hasta ser ultrainteligente si no ha desarrollado antes una personalidad benevolente.

Numerosas figuras del mundo de la ciencia y la tecnología, como Stephen Hawking, el premio Nobel de Física Frank Wilczek o los empresarios Bill Gates, Steve Wozniak, cofundador de Apple, y Elon Musk, fundador de Tesla y SpaceX, han expresado reticencias similares a las formuladas por Bostrom. En 2014 se unieron a otros cientos de expertos y firmaron un documento en el que recomendaban la investigación en inteligencia artificial, pero advirtiendo al mismo tiempo de que estos sistemas «deben hacer lo que nosotros queramos que hagan», sobre todo en el ámbito de armas autónomas que puedan actuar sin intervención humana.

Otros expertos, por el contrario, conducen sus argumentos por terrenos menos distópicos y sostienen que una IA lo bastante sofisticada podría ser capaz de detectar fallos en su propio diseño y modificarse a sí misma para ser segura. El propio Hawking, además de advertir de los peligros de la IA, admite también que su desarrollo convenientemente supervisado podría ofrecer a la especie humana una nueva era de abundancia y prosperidad. Pero ¿hasta qué punto es plausible el escenario de la singularidad y de la ultrainteligencia que nos conduciría a ella? Para responder a esta pregunta, primero hay que determinar cuán cerca estamos de construir una inteligencia artificial similar a la humana, o si se trata de algo siquiera posible.

LA PROBABILIDAD DE UNA INTELIGENCIA NO HUMANA

La mayor parte de los científicos, incluso aquellos cuya actividad pueda estar muy alejada de la computación o la neurociencia, estarían seguramente de acuerdo en que no hay nada en la naturaleza del cerebro que suponga un freno insuperable a la consecución de una IA. Los cerebros, al fin y al cabo, son máquinas biológicas que

se rigen por las leyes básicas de la física; todo en ellos, por tanto, ha de poderse analizar y simular. En el peor de los casos, se aduce, si un proceso ciego como la evolución ha sido capaz de producir la inteligencia humana, ese mismo proceso, adecuadamente optimizado gracias a la ciencia, podrá hacerlo nuevamente y de forma más rápida. Bajo esta perspectiva, la IA se convierte en un problema fundamentalmente tecnológico donde la cuestión relevante ya no es el *si*, sino más bien el *cómo* y, en menor medida, el *cuándo*.

Una encuesta publicada en 2016 preguntaba a los 100 investigadores más citados en el ámbito de la IA en qué año creían que se lograría una IA de nivel humano. El 50% pensaba que antes de 2050, y este porcentaje subía al 90% si la fecha se retrasaba a 2070. Además, un 79% de estos mismos expertos opinaba que las consecuencias de la IA serían positivas o neutras. Aun en el supuesto de que la esperanza de vida se mantenga en sus valores actuales, un europeo que tenga hoy diez años casi seguro será testigo, según esa encuesta, del nacimiento de una inteligencia no humana.

En esta misma encuesta, los expertos respondieron también a la pregunta sobre qué línea de investigación era la que iba a contribuir más al objetivo de construir una IA. Estas fueron las diez candidatas con más votos:

Neurociencia computacional	42%
Arquitecturas cognitivas	42%
Redes neuronales	40%
Mayor poder de computación	37%
Bases de datos masivas	35,5%
Mentes corporeizadas	35%
Otros métodos desconocidos a día de hoy	32,5%
Simulación cerebral completa	29%
Algoritmos evolutivos	29%
Sistemas lógicos	22%

La neurociencia computacional, primera línea de investigación fructífera, se ocupa de estudiar hasta qué punto el procesado de información del cerebro es susceptible de ser modelado computacionalmente. Su valor a efectos del objetivo de la IA estriba en proporcionar modelos de cognición humana traducibles a programas de ordenador. Se trata de una disciplina nacida a finales del siglo pasado que todavía ha de producir un gran resultado, pero su potencial, como refleja la encuesta, es enorme. Algunos hallazgos procedentes de la neurociencia computacional tienen ya reflejo en sistemas de IA funcionales, como las redes neuronales artificiales de Hopfield, que emulan la memoria asociativa humana, o una amplia variedad de modelos de reconocimiento visual y táctil implementados en robots semiinteligentes.

El término *arquitectura cognitiva*, por su parte, se refiere a programas que simulan la cognición humana. Las redes neuronales serían un caso. Con *arquitectura cognitiva mixta* nos referimos a programas que combinan tecnologías de varios tipos. Una arquitectura de esta categoría podría compaginar, por ejemplo, una red neuronal que se encargara de simular el aprendizaje y un sistema experto que hiciera lo propio con la toma de decisiones. Llegados a este punto, tal vez convendría detenerse un poco en las redes neuronales, vista su ubicuidad entre las líneas de investigación más prometedoras.

Una red de neuronas artificiales conectadas entre sí

Las redes neuronales artificiales, o RNA, suponen un paradigma de programación completamente diferente de la computación tradicional. Una RNA no sigue un patrón lineal de cálculo, sino que procesa la información de forma colectiva a través de una red de neuronas artificiales que emula el funcionamiento del cerebro humano.

El funcionamiento de la neurona artificial dentro de estas redes es muy sencillo: procesa la información que le llega de otras neuro-

nas conectadas a ella y produce un resultado. Sin embargo, un conjunto de neuronas, adecuadamente conectadas, puede reproducir un comportamiento muy inteligente. Las neuronas en una RNA están dispuestas en capas. La primera capa se denomina *capa de entrada,* y en ella se codifica el problema que se quiere que la RNA resuelva. Esa capa conecta con otras neuronas dispuestas en una o más *capas intermedias,* que a su vez conectan con una capa final o *capa de salida,* que emite el resultado. Una neurona solo transmite información si la suma de los valores que le llegan supera un determinado valor de referencia o umbral. En caso contrario, la neurona inhibe la señal. La combinación de valores necesaria para superar el umbral de una neurona y que esta propague la señal es uno de los elementos fundamentales del diseño de una RNA.

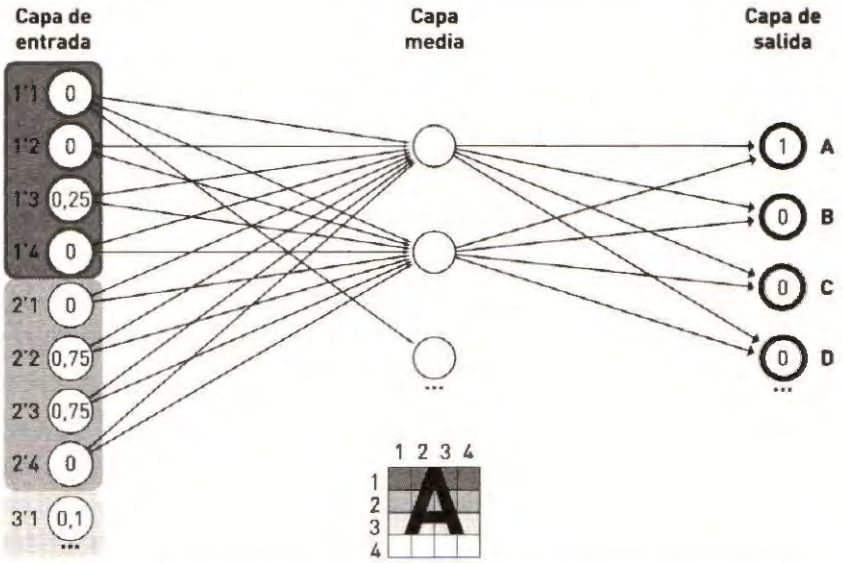
Un ejemplo, como el ilustrado en la figura 1 de la página 25, nos ayudará a entender mejor la teoría. Imaginemos que queremos identificar caracteres manuscritos, por ejemplo, una letra. Lo primero sería descomponer la letra en píxeles, y representar cada píxel por una neurona de entrada. Esa información se trasladaría a las neuronas de la capa intermedia, que, en función de los umbrales y de las reglas con las que se haya programado la red, propagarán la señal hacia una capa intermedia adicional o directamente a la capa de salida. El resultado en la capa de salida representaría la solución al problema de cuál es el carácter en cuestión.

En nuestro ejemplo, queremos ver si la red neuronal es capaz de identificar una letra A. La capa de entrada de la red neuronal está compuesta por 16 neuronas, una por cada uno de los píxeles que emplearemos en representar la imagen de la letra. Para ello, superpondremos sobre el carácter una cuadrícula de 4x4. Cada neurona de la capa de entrada representa una celda de la cuadrícula y adopta un valor numérico del 0 al 1, en el que 0 indica que la celda está vacía, 1 que está totalmente llena, y un valor comprendido entre 0 y 1, que la celda está llena en ese porcentaje. En lo que respecta a la capa de salida, hemos dispuesto una neurona por cada carácter del alfabeto.

Al arrancar la RNA, los valores de la capa de entrada se propagan hacia adelante, y siguen propagándose en función de si superan o no los umbrales determinados por el diseño. De este mecanismo acaba resultando un valor para cada neurona de la capa de salida. Si la que representa la letra A tiene un valor de 1 significa que la red ha predicho el carácter A para ese carácter de entrada sin ningún tipo de duda (lo que en nuestro ejemplo resultaría un éxito). Podría ser también que en lugar de una única neurona de valor 1 varias neuronas tuvieran números entre 0 y 1. Una RNA siempre tiene una proporción de incertidumbre o error. Por ejemplo, podría darse el caso de que el valor de la neurona que representa la D fuera de 0,80, y el de la neurona que representa la O de 0,70. En este caso, la RNA estaría prediciendo que el carácter de entrada es, bastante probablemente una D o, de forma solo un poco menos segura, una O. Es lo mismo que nos pasa a cualquiera de nosotros cuando leemos un texto manuscrito: si la caligrafía no es clara, podemos tener dudas sobre alguna letra.

Antes de poner en marcha la RNA, sin embargo, es necesario entrenarla. Esto significa que la RNA tiene que aprender qué reglas aplicar para propagar o no una señal en un caso concreto. En el ejemplo tratado, el programador recurre a caracteres de los que ya se sabe la letra que representan. Si, una vez introducidos en la capa de entrada, la RNA los identifica correctamente, es porque las reglas son las adecuadas. En caso contrario, hay que alterarlas en mayor o menor grado para obtener el resultado correcto. Esta corrección a partir de si el dato de salida es o no correcto, es decir, hacia atrás, se produce de forma automática. A medida que introducimos más y más datos de entrenamiento e indicamos a la RNA cuáles son correctos y cuáles no, la RNA se va ajustando de forma cada vez más precisa. Por tanto, cuantos más y mejores datos, mayor será la capacidad predictiva de una RNA. Este proceso de entrenamiento o aprendizaje se repite hasta obtener un error de predicción aceptable, momento en el que ya pueden introducirse caracteres no identificados para que la RNA los identifique.

Fig. 1



Las RNA modernas se componen de miles de neuronas artificiales dispuestas en varias capas: una primera, de entrada, que codifica el problema; una o más capas intermedias, donde se genera la solución; y una capa de salida, que la recoge.

Las RNA llevan a cabo tareas que son extremadamente fáciles para un humano, como reconocer imágenes, pero que históricamente las máquinas eran incapaces de gestionar a causa del enorme poder de computación requerido. Aunque sus fundamentos teóricos se establecieron en la década de 1950, ha sido solo recientemente cuando han desplegado todo su potencial gracias a la confluencia de la creciente capacidad de computación de los ordenadores actuales con la enorme cantidad de datos disponibles en internet. Precisamente, la existencia de bases de datos de gran tamaño, una de cuyas utilidades es entrenar a RNA cada vez más sofisticadas, es otro de los factores que han de desempeñar un papel importante en el advenimiento de las IA, según un 35,5% de los encuestados.

Un buen ejemplo del poder del aprendizaje automático basado en las RNA es el de ImagenNET, un concurso anual que premia a la red neuronal artificial que clasifique mejor un grupo de imágenes. La ganadora de 2011 clasificó imágenes con una tasa de error del 25,8%. En 2012, la tasa de error era solo del 16,4%. En 2013 fue del 11,7%. En 2014, del 6,7%. En enero de 2015 se logró una tasa de error en el reconocimiento de imágenes del 6%. En febrero, Microsoft alcanzó el 4,9%. En marzo, Google llegó al 4,8%. El ser humano tiene una tasa de error que ronda el 5%. Es decir, las RNA ahora ya son capaces de reconocer imágenes mejor que un ser humano.

Las RNA están concebidas, básicamente, para reconocer patrones allá donde los datos las muestren. Sus aplicaciones son, por tanto, amplísimas. Tanto AlphaGo como Watson, las IA que saltaron a la palestra en su día tras vencer a contrincantes humanos en el go y en el concurso Jeopardy! respectivamente, integran el aprendizaje por RNA entre sus componentes. Y apuntemos una aplicación más: la lectura de labios. Si una persona experta en lectura de labios es capaz de acertar alrededor de un 50% de las veces, las RNA alcanzan, en el momento de redactar este libro, el 93,4%. Se ha escogido este ejemplo por dos razones: la primera, por la notable ventaja de las RNA con respecto al estándar humano. ¿Y la segunda? Porque es leyendo los labios del astronauta Bowman como HAL 9000, la IA protagonista de la célebre *2001: Una odisea del espacio*, adivina los planes de aquel para desconectarla.

El argumento del desarrollo tecnológico

Hasta ahora hemos visto adelantos en lo que podría denominarse el *software* de la IA. Ahora bien, si las inteligencias no humanas han de ser una realidad en nuestra vida, o al menos en la de nuestros hijos, ¿hasta qué punto han de darse también avances en el *hardware*, es decir, en los ordenadores? Ello dependerá de cuán exigentes

en poder de computación sean las tecnologías que acaben siendo necesarias, lo que resulta una incógnita a día de hoy. Pero con independencia de ese dato, un factor induce al optimismo: el llamado *crecimiento exponencial de la tecnología*. Se trata de un concepto estrechamente relacionado con la llamada *ley de Moore*. Esta ley mantiene que cada veinticuatro meses, aproximadamente, la potencia informática crece de manera exponencial, porque el número de transistores por pulgada en los circuitos integrados se duplica. Esta constante, que se ha venido produciendo desde que en 1965 la formulara el cofundador de Intel, Gordon Moore, ha permitido que en poco más de medio siglo el chip de 64 transistores disponible aquel año disponga de miles de millones de transistores en la actualidad.

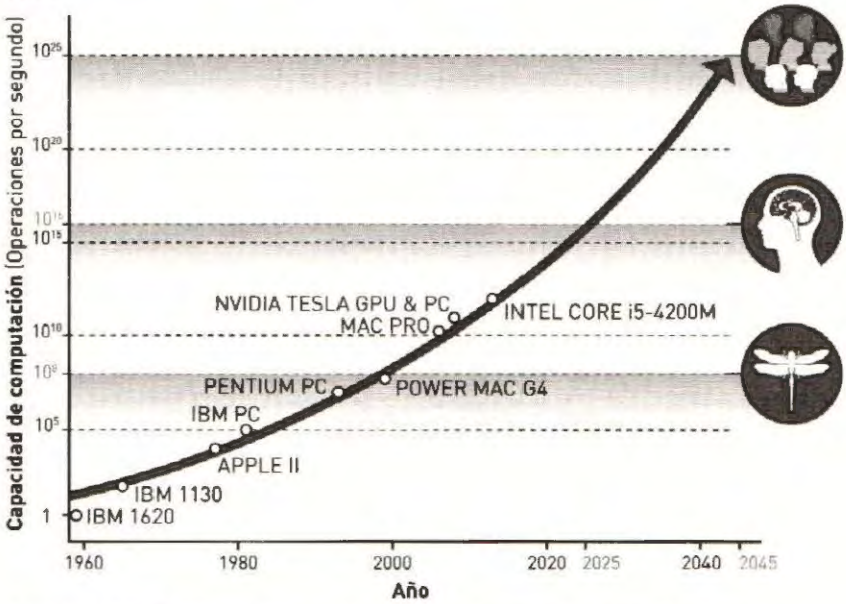
Estas cifras son relativamente abstractas, pero lo destacable no es tanto la cifra de transistores en sí, como su acelerado ritmo de crecimiento. En cualquier caso, resulta muy difícil imaginar esta progresión y, sobre todo, advertir cuán rápido se alcanzan cifras extraordinariamente elevadas. Por ello, para asimilar el comportamiento del crecimiento exponencial se suelen emplear analogías y metáforas más o menos aproximadas. La más famosa de ellas es la del tablero de ajedrez y los granos de trigo. Cuenta la leyenda cómo el supuesto inventor del ajedrez, un matemático indio, le mostró un día su hallazgo al rey de un lejano país de Oriente, quien quedó tan impresionado que le ofreció la recompensa que él mismo dispusiera. El sabio no se hizo de rogar y presentó una petición que, en un primer momento, pareció ridículamente modesta al monarca: un grano de trigo por la primera casilla o escaque del tablero de ajedrez, dos granos por la segunda casilla, cuatro por la tercera, y así sucesivamente hasta llegar a la última casilla, la número 64. El rey creía que bastaría con un buen saco de trigo como pago, pero solo la casilla 32, en la mitad del tablero, suponía ya el desembolso de 4 000 millones de granos trigo. Al alcanzar la casilla 64 se puso de manifiesto la magia del crecimiento exponencial: se requerían más de 18 trillones de granos

de trigo, más que todo el trigo que había en el mundo, para pagar al sabio. De hecho, serían necesarias las cosechas mundiales de más de 22000 años para satisfacer semejante petición.

Todavía más sorprendentes resultan los efectos del crecimiento exponencial si comparamos la distancia que recorreremos dando pasos lineales y pasos exponenciales. Por ejemplo, si damos treinta pasos a razón de un metro por paso, habremos cubierto treinta metros. Pero si damos treinta pasos exponenciales (el primero de un metro, el segundo de dos metros, el tercero de cuatro metros, ocho, dieciséis, treinta y dos...), al llegar a los treinta pasos habremos dado veintiséis vueltas a la Tierra. En sus primeras duplicaciones, el crecimiento exponencial resulta muy engañoso. Estas son irrisorias, y de ahí pasan a ser más bien poco llamativas, casi indistinguibles del crecimiento lineal: 2, 4, 8, 16, 32, 64, 128, 256, 512... Tras nueve duplicaciones, continuamos barajando cifras no demasiado espectaculares. Sin embargo, llegados a cierto punto, cada duplicación supone aumentar la cifra en una proporción extraordinariamente mayor a la anterior. Ese punto en el que el crecimiento exponencial se dispara ha sido bautizado como *inflexión de la curva*. Y es precisamente en esta inflexión de la curva donde se halla la capacidad de computación actual de los ordenadores.

Observemos el caso concreto del número de operaciones por segundo que era capaz de realizar un ordenador portátil de precio medio a principios de la década de 2000: una media de 10^8, es decir, un uno seguido de ocho ceros (fig. 2). Este es aproximadamente el número equivalente de operaciones por segundo que lleva a cabo el cerebro de un insecto. Gracias al crecimiento exponencial, cinco años después, un ordenador similar era capaz de realizar 10^{11} operaciones por segundo, es decir, el equivalente aproximado al cerebro de un ratón. Si continuase operando el crecimiento exponencial del número de transistores por pulgada en un circuito integrado, en el año 2025 un ordenador como el que todos podemos tener en casa dispondrá de la misma capacidad de computa-

Fig. 2



La ley de Moore explica que la capacidad de computación de los sistemas informáticos crece de manera exponencial cada año. Teniendo en cuenta esta constante, en 2025 un ordenador tendrá la capacidad de cálculo de un ser humano; y en 2045, la de toda la humanidad.

ción que un cerebro humano: 10^{16} operaciones por segundo. Si se continúa al mismo ritmo de crecimiento, en 2045 un ordenador doméstico tendrá la misma capacidad de computación que todo el conjunto de la humanidad (se prevé que unos 9 000 millones de personas). Concebir lo que supondría esta potencia de cálculo es algo que excede a nuestra imaginación, y más aún si se tiene en cuenta que no habrá un solo ordenador con semejante capacidad, sino millones de ellos.

Ahora bien, la ley de Moore tiene límites físicos infranqueables, como su propio creador reconoció en 2007, cuando dijo que su ley dejaría de cumplirse en unos 15 años. Esto es así porque, llegada a

cierto punto, la miniaturización de un transistor (el núcleo del microprocesador) provoca que este deje de funcionar correctamente. Hasta el momento se ha logrado un nivel de miniaturización que permite que cada transistor esté inscrito sobre 14 nanómetros. Se pronostica que podríamos alcanzar de 5 a 7 nanómetros; pero, a partir de ese punto, los electrones se escapan de los canales por donde deben circular. Para asimilar este grado de miniaturización y los desafíos técnicos que supone, hemos de tener en cuenta que un simple virus tiene un diámetro de 75 nanómetros; la pared de una bacteria, 10 nanómetros; y una molécula de glucosa, 1 nanómetro.

> En el siglo XXI no experimentaremos 100 años de progreso, sino 20000 (al ritmo de hoy).
>
> RAY KURZWEIL

Algunos defensores del crecimiento exponencial consideran que, si bien la ley de Moore es finita, siempre podremos encontrar tecnologías sustitutivas cuyo horizonte de crecimiento sea más amplio. Por ejemplo, durante la primera Revolución Industrial, entre 1750 y 1830, la tecnología vigente, basada en la energía del vapor, alcanzó el límite de su eficiencia y solo el desarrollo de una tecnología alternativa, la electricidad o el motor de combustión, permitió superarlo, dando lugar a la segunda Revolución Industrial, entre 1870 y 1914. Estos cambios de paradigma, en lo sucesivo, se producirán cada vez más deprisa. Uno de estos nuevos paradigmas podría ser la computación cuántica.

La base de la tecnología digital son los transistores, dispositivos electrónicos binarios que alternan entre dos estados: 0 o 1, verdadero o falso. Un único transistor es capaz de codificar solo dos valores (o un *bit*) : 0 o 1. Dos transistores, por su parte, pueden codificar ya cuatro valores (dos bits): 00, 01, 10 u 11. Con cuatro transistores, los valores codificados son ya 16 (8 bits); con 20, 2^{20} valores (es decir, 1.048.576), y así sucesivamente. En general, n transistores pueden codificar 2^n valores.

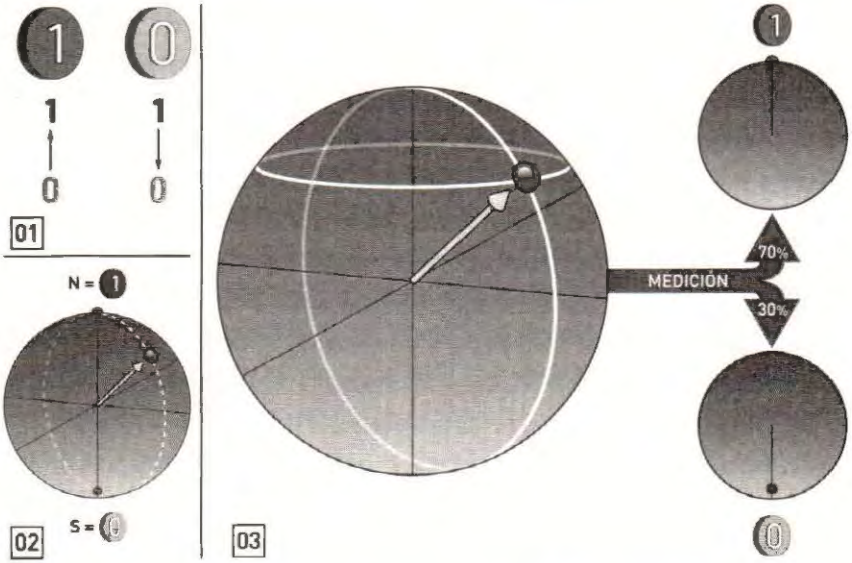
Para hacer cálculos, los chips de los ordenadores combinan miles de millones de transistores con los que ejecutan operaciones lógicas sencillas. La potencia computacional, por tanto, es mayor cuantos más transistores contenga un chip y, a la vez, tiene como límite máximo el número de transistores que se pueden implementar en un chip. Como ya hemos dicho, dicho límite se prevé que se alcance en 2025.

La computación cuántica permitiría franquear esta barrera. Las unidades mínimas de almacenamiento en computación cuántica se denominan *qubits*. Ya hemos visto que un bit solo puede almacenar dos valores, 1 o 0. Los qubits, en cambio, almacenan esos dos valores binarios más cualquier superposición cuántica de 0 y 1 (fig. 3). De esta forma, un qubit puede codificar al mismo tiempo 0 y 1, es decir que admite dos valores; un par de qubits pueden codificar al mismo tiempo 4 estados o valores diferentes; tres qubits pueden codificar al mismo tiempo 8 estados diferentes, etc. En general, un ordenador cuántico con n qubits puede estar en una superposición de hasta 2^n estados diferentes *simultáneamente*. Un ordenador binario con n bits, en cambio, puede estar en uno solo de esos 2^n estados diferentes en un momento dado. En un periodo de tiempo dado, por tanto, el ordenador cuántico ha sido capaz de ejecutar un número increíblemente mayor de operaciones de forma paralela. La computación cuántica es, pues, mucho más que un medio para salvar el límite físico de la ley de Moore: es un paradigma de computación radicalmente nuevo, con una potencia de cálculo mucho mayor que la computación binaria clásica.

Otros caminos posibles a la IA humana

Con todo, a pesar de que confiemos en las nuevas tecnologías emergentes de crecimiento exponencial, donde tecnologías revolucionarias tomarían el relevo de las tradicionales en el caso de agotarse

Fig. 3



Un bit puede tener dos estados, 1 o 0, representados de forma abstracta por una flecha arriba o abajo (**01**). Un qubit, por su parte, puede tener infinitos estados, representados por una flecha que apunta a una localización en la superficie de una esfera, con los valores 1 y 0 en los polos norte y sur, respectivamente (**02**). Cuando se mide el estado del qubit, este se colapsa y adopta un valor binario, 0 o 1, igual que un bit clásico. La probabilidad de uno u otro valor depende de la localización (**03**).

estas últimas, lo cierto es que históricamente no suelen darse tasas de crecimiento tan altas fuera del ámbito de la computación. De hecho, la mayoría de los avances científicos han seguido un patrón irregular y no exponencial.

Si finalmente la IA topara con algún obstáculo tecnológico o biológico infranqueable, ello no supondría el bloqueo total del camino a la ultrainteligencia. Al menos uno de esos caminos podría seguir siendo viable: la simulación cerebral completa. Aunque la encuesta lo incluyó entre las tecnologías propias de la IA (y le asignó un 29% de probabilidades de ser un factor decisivo), se diferencia de aque-

llas en que no hay un intento de reducir la cognición a una serie de reglas o de generarla, como en las RNA, a partir de componentes artificiales de inspiración biológica; no, de lo que se trata aquí es de un plagio directo. En comparación con la IA tradicional, se basa menos en el conocimiento teórico y más en la capacidad tecnológica.

La tarea de simular con éxito un cerebro humano completo tendría tres grandes fases: el escaneo del original biológico, la reconstrucción tridimensional en un ordenador y la implementación de la copia. Es importante tener en cuenta que el éxito de la simulación de cerebro completo no radica principalmente en la fidelidad de la copia, sino en que dicha copia capture las propiedades computacionalmente funcionales del cerebro. Es decir, que la copia sea un programa que, una vez ejecutado, sea capaz de pensar como un cerebro.

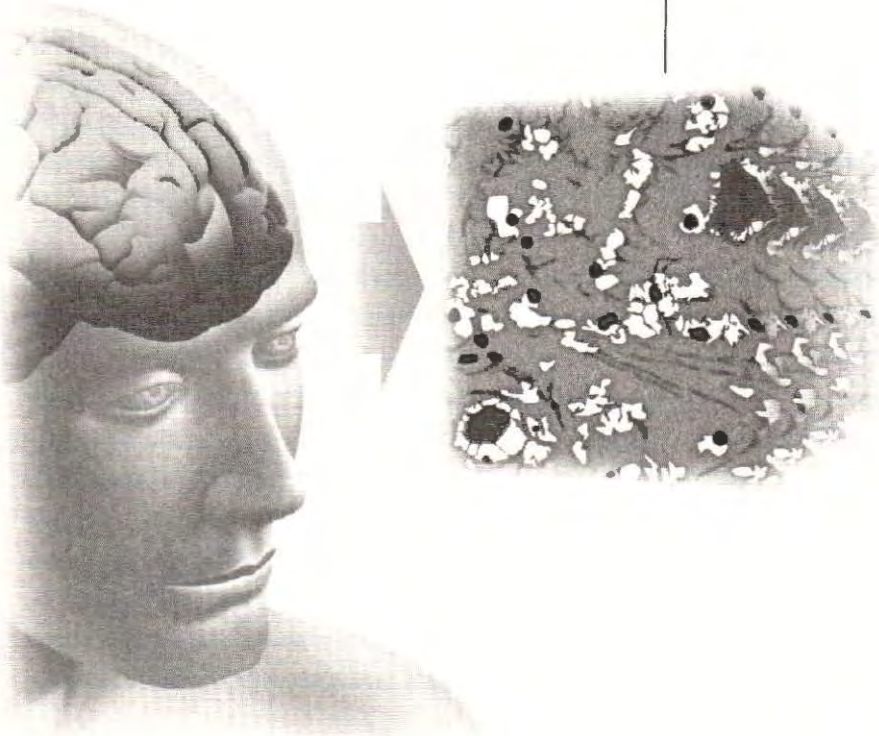
Con todo, sigue suponiendo un reto extraordinario. Modelos extremadamente simples de neuronas individuales requieren alrededor de 1000 operaciones por segundo. Se estima que, para simular un cerebro completo a nivel molecular pueden necesitarse hasta 10^{43} operaciones por segundo; un supercomputador actual proporciona del orden de 10^{16} operaciones por segundo. Todavía no se ha logrado una simulación funcional de un cerebro más sencillo que el de un humano, como pueda serlo, por ejemplo, el del *Caenorhabditis elegans*, un gusano de aproximadamente 1 milímetro de longitud que solo tiene 302 neuronas y 5500 conexiones sinápticas (el cerebro humano tiene entre 50000 y 100000 millones de neuronas y, por lo menos, 10^{14} conexiones sinápticas). De hecho, ni siquiera se ha completado la simulación previa del cuerpo, músculos y entorno en los que interactúa este gusano para, después, pasar a un modelo que simule el comportamiento y la actividad neuronal de su cerebro.

Además de un mayor poder de computación, o sea, un *hardware* lo suficientemente potente como para crear un modelo completo del cerebro, también se necesitará un mayor grado de detalle en el escrutinio de su funcionamiento. Nos referimos aquí a tecnologías

> LA EMULACIÓN CEREBRAL COMPLETA

La hoja de ruta de una simulación del cerebro humano se compone de tres fases principales: el escaneado del órgano, su traducción a un modelo computacional que sea exacto tanto en lo fisiológico como en lo funcional y, finalmente, la simulación propiamente dicha.

Fase 1: Escaneado
En esta etapa, el objetivo es capturar el cerebro con detalle suficiente como para poderlo reconstruir con posterioridad. El reto tecnológico principal se da en la microscopía.

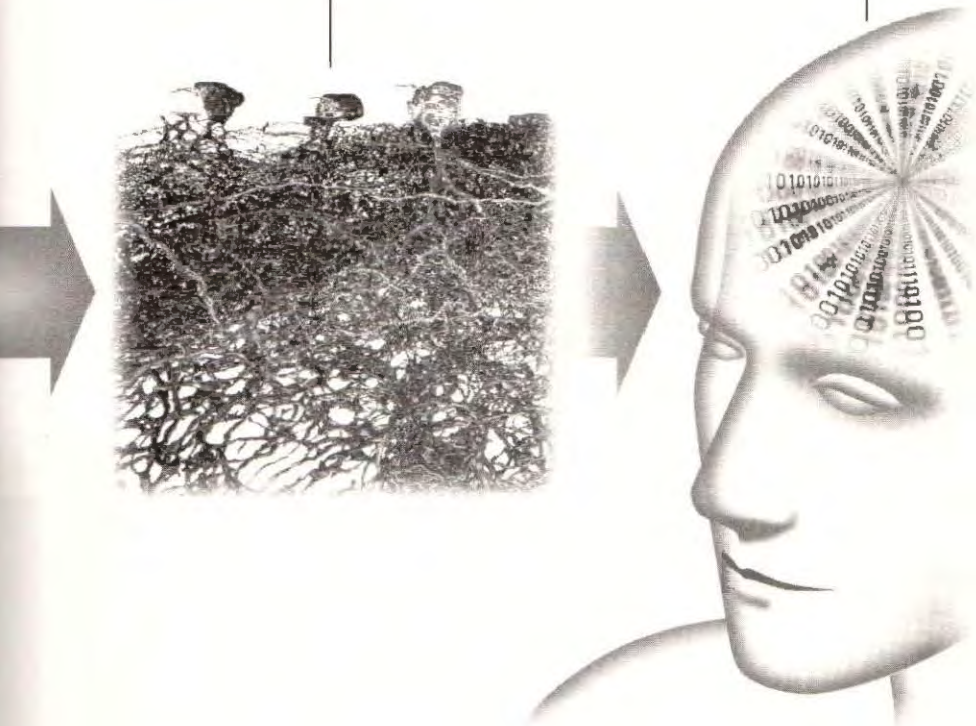


Fase 3. Simulación

En esta etapa final, la estructura neurocomputacional resultante se ejecuta en un ordenador. Es muy posible que deban simularse también un cuerpo físico y un entorno con el que interaccionar.

Fase 2. Traducción

Los datos en bruto del escaneo deben pasarse por un programa de reconocimiento de imagen que identifique los elementos relevantes y les restituya el volumen original. A esta réplica se la tiene que acompañar de modelos computacionales de funcionamiento.



que intervienen en la primera fase, la del escaneado: microscopios más potentes capaces de detectar propiedades relevantes del cerebro, por un lado, o instrumentos de análisis automático de las imágenes obtenidas que transforme estos datos brutos en modelos.

Una vez completado ese escaneado, y establecido un diagrama de conexiones de todo el cerebro, se debería cuantificar con gran exactitud cómo interactúan las neuronas en cada punto de unión a nivel molecular. Para hacernos una idea de la complejidad que ello entraña, ni siquiera conocemos el número de moléculas que hay en un cerebro y mucho menos las que son importantes para su funcionamiento. Así pues, solo podremos extraer las operaciones digitales y lógicas de un cerebro si conocemos en profundidad todos estos elementos y cómo interactúan entre ellos con un nivel de precisión aún inimaginable.

De todos modos, podría no ser necesario apurar todo este proceso hasta sus últimas consecuencias. Porque no es descabellado suponer que, en algún punto del proceso que haya de conducir a la simulación completa, los adelantos acumulados hasta el momento en forma de emulaciones parciales confluyan en una IA.

ESCENARIOS POSIBLES TRAS UNA EXPLOSIÓN DE INTELIGENCIA

Suponiendo que la singularidad llegue, ¿qué supondría para la humanidad? ¿Qué posibles escenarios podríamos encontrarnos? El filósofo David Chalmers, en un influyente artículo de 2010, aventuró cuatro: extinción, aislamiento, inferioridad e integración. La posibilidad de que la ultrainteligencia suponga un peligro existencial ya ha sido explorada al principio del capítulo. El aislamiento nos evitaría ese riesgo, pero implicaría desaprovechar el enorme potencial de la ultrainteligencia. La coexistencia en inferioridad podría ser el único modo de disfrutar de ciertos adelantos tecno-

lógicos. No obstante, es seguramente preferible un escenario en el que podamos considerarnos los iguales de la IA. La opción de la integración quizá implicará abandonar, gradualmente o de una vez, nuestras raíces biológicas. La opción más plausible sería alguna forma de migración o subida de la mente individual a un entorno computacional, un proceso semejante a la simulación cerebral completa del apartado anterior. Sin embargo, esta posibilidad abre nuevas incógnitas. Esa versión computacional de nuestras mentes ¿compartiría con la versión biológica todo aquello que nos hace *nosotros*? ¿Una conciencia, una identidad?

Como veremos en el capítulo siguiente, el padre de la IA, Alan Turing, ya se planteó preguntas parecidas. La diferencia es que, en este futuro hipotético que estamos explorando, los objetos de escrutinio ya no serían las máquinas, sino los propios seres humanos. ¿O ya no cabrá tal distinción? Porque el resultado final de la integración sería convertirnos, nosotros mismos, en ultrainteligencias.

02 DE TURING AL *BIG DATA*

Tras un inicio arrollador, el campo de la IA acabó enterrado bajo el peso de las expectativas generadas. Hoy, con el *big data* alimentando tecnologías como las redes neuronales, el objetivo de alcanzar una inteligencia equiparable o superior a la humana vuelve a ser plausible.

La talla del matemático británico Alan Mathison Turing no ha hecho sino crecer desde su prematura muerte en 1954, a medida que el ordenador influye cada vez más en la vida y la sociedad contemporáneas. Si alguna figura mereciera el apelativo de «padre de la computación», ese sería el británico (con el permiso, quizá, de John von Neumann). A Turing se deben contribuciones fundamentales en varios ámbitos clave: la teoría de la computación, la arquitectura de ordenadores y la criptografía. Sin embargo, en este libro vamos a centrarnos únicamente en sus aportaciones a un cuarto ámbito: el de la inteligencia artificial.

Definir de forma precisa qué es la inteligencia o qué caracteriza a una entidad inteligente es una labor ardua sobre la que todavía no existe un acuerdo mayoritario. La definición que propuso Turing fue una de las primeras, y también una de las más influyentes. El británico publicó su idea en un artículo de 1950 y se valió para explicarla de un juego de sociedad popular en la Inglaterra victoriana, el juego de la imitación: consistía en encerrar a un hombre y a una mujer en sendas habitaciones y obligarles a responder a las

preguntas de un tercero siempre con verdades, el hombre, o siempre con mentiras, la mujer. El encuestador debía entonces adivinar en qué habitación se escondía cada cual. En la variante propuesta por Turing, los actores encerrados en las habitaciones pasaban a ser una máquina y una persona, y el juego consistía en identificar correctamente al ser humano (fig. 1). Si la máquina conseguía engañar al encuestador, había pasado el test y, según Turing, había demostrado inteligencia.

El *test de Turing*, como esa prueba ha pasado a conocerse, propone una definición puramente operativa de inteligencia: es inteligente aquello que se comporta de forma inteligente. A Turing solo le preocupa lo que se percibe desde fuera de la habitación. Lo que pase dentro de ella, es decir, el modo en el que la máquina hace la imitación, es irrelevante a efectos de decidir si es o no inteligente. La mayoría de las críticas al test de Turing le reprochan, precisamente, este olvido. Lo que ocurre dentro de la habitación, vienen a decir, importa y mucho.

Con el ánimo de rebatir la idea de inteligencia de Turing, el filósofo estadounidense John Searle concibió un experimento mental conocido como *la habitación china*. El ejercicio nos invita a imaginar una persona encerrada en una habitación. Alguien situado en el exterior introduce en la estancia una serie de preguntas en chino a través de una bandeja de entrada. La persona de la habitación debe generar una respuesta a los mensajes y depositarla en una bandeja de salida. Supongamos que, una vez leídas las respuestas, el interrogador es incapaz de diferenciar si en la habitación hay una persona o una máquina. La persona que está en la habitación, sin embargo, ha hecho lo siguiente: consultar un manual de chino donde se indica qué respuesta debe dar a cada mensaje. Según Searle, no hay diferencia real entre una máquina y el ocupante de la habitación china a efectos del test de Turing; ambos han demostrado ser inteligentes. Sin embargo, el individuo de la habitación ¡no ha entendido absolutamente nada de los mensajes!

Fig. 1

En el test de Turing, un interrogador, sentado en la estancia central, conversa a través de las pantallas con una máquina, en la habitación izquierda, y una persona, en la habitación derecha. Unos jueces vigilan el proceso. Si la máquina les hace creer que es humana, supera la prueba.

Aunque el test de Turing es la más destacada de las aportaciones del matemático a la inteligencia artificial, no es la única. A él se deben también importantes e influyentes reflexiones acerca del mejor modo de abordar la tarea de construir una inteligencia no humana. En 1950 escribía:

> En lugar de tratar de producir un programa que simule la mente adulta, ¿por qué no intentar producir uno que simule la de un niño? Si la sometiéramos entonces a una educación apropiada, obtendríamos el cerebro adulto.

En este simple párrafo se encuentra el germen de uno de los paradigmas de la inteligencia artificial contemporánea, el basado en el aprendizaje. Sobre el potencial de esta máquina infantil, Turing añadió:

> No podemos esperar encontrar una buena máquina infantil al primer intento. Debemos experimentar enseñando a una de estas máquinas y ver cómo aprende. Luego, podemos probar con otra y

> comprobar si es mejor o peor. Existe una relación obvia entre este proceso y la evolución. [...] Podemos esperar, no obstante, que este proceso sea más rápido que la evolución. La supervivencia del más apto es un método lento para medir ventajas. El experimentador, gracias a su inteligencia, debería ser capaz de acelerarlo.

Es precisamente el poder de los procesos evolutivos dirigidos, como el señalado por Turing, el que llevó a destacados teóricos como Hans Moravec o David Chalmers a defender la viabilidad no solo de generar una inteligencia humana artificial, sino de hacerlo en el siglo XXI. Al fin y al cabo, si la evolución ha sido capaz de producir inteligencia, ¿cómo no va a producirla, y a una mayor velocidad, la evolución acelerada por la ingeniería humana?

IA EN DOS DIRECCIONES: SIMBOLISTAS CONTRA CONEXIONISTAS

Después del punto de inflexión marcado por Turing, tanto los medios de comunicación como los pensadores de la época empezaron a plantearse seriamente la posibilidad de que las máquinas llegaran a sustituir al ser humano. La revista *Fortune*, por ejemplo, llevó a su portada el titular *Máquinas sin hombres* (1946), y el filósofo británico Bertrand Russell se preguntó en 1951: «¿Son los humanos necesarios?» Con todo, fue en una conferencia celebrada en 1956 en la Universidad de Dartmouth, en New Hampshire, organizada por los informáticos y expertos en ciencias cognitivas John McCarthy y Marvin Minsky, donde nació de facto el campo de investigación de la Inteligencia Artificial moderna, es decir, el propósito real y factible de concebir una máquina pensante. En la declaración fundacional del encuentro, al que acudieron, entre otros, Claude Shannon, el creador de la moderna teoría de la información, y Herbert A. Simon, futuro premio Nobel de Economía, se estableció:

> ALAN TURING, EL PRECURSOR DE LA INFORMÁTICA MODERNA

Ya de joven, Alan Turing demostró notables aptitudes para las matemáticas. En 1936, cuando contaba apenas con 24 años, ideó una máquina hipotética capaz de simular cualquier operación lógica expresada en forma de algoritmo. La *máquina de Turing* acabó siendo el modelo teórico sobre el que se diseñaron las unidades de procesamiento (CPU) de los ordenadores modernos. Con el estallido de la Segunda Guerra Mundial, el matemático se incorporó al equipo de analistas criptográficos del Gobierno británico, situado en las instalaciones militares de Bletchley Park. Allí contribuyó a descifrar los códigos nazis emitidos por la máquina Enigma, con lo que se calcula que contribuyó a salvar más de diez millones de vidas. En 1948, entró a trabajar en la Universidad de Manchester, donde desempeñó un papel destacado en la construcción de los primeros ordenadores con almacenamiento de memoria o de los transistores. En 1950 hizo pública la idea de su célebre test, y dos años más tarde, fue detenido y condenado por mantener relaciones impropias con un joven de 19 años. Tras aceptar someterse a una terapia química que supuestamente reducía su deseo sexual, Turing se suicidó el 7 de junio de 1954 con una dosis de cianuro, probablemente introducida en una manzana.

— Alan Turing en 1951, el punto álgido de una carrera que se truncaría un año después a causa de su condena por homosexualidad.

> El estudio es para proceder sobre la base de la conjetura de que cada aspecto del aprendizaje o cualquier otra característica de la inteligencia permiten, en principio, ser descritos con tanta precisión que puede fabricarse una máquina para simularlos. Se intentará averiguar cómo fabricar máquinas que utilicen el lenguaje, formen abstracciones y conceptos, resuelvan las clases de problemas ahora reservados para los seres humanos, y mejoren por sí mismas.

Es decir, máquinas que razonen y comprendan, se comuniquen y aprendan. La declaración terminaba con estas palabras:

> Creemos que puede llevarse a cabo un avance significativo en uno o más de estos problemas si un grupo de científicos cuidadosamente seleccionados trabajan en ello conjuntamente.

Este optimismo tecnológico era de esperar en una época en la que el hombre había demostrado ser capaz de domesticar el átomo. Pero la curiosidad científica no era el único combustible que alimentaba las ambiciones del programa de Dartmouth. No debemos olvidar que en esa década EE. UU. y la URSS iniciaban un conflicto silencioso y frío que, a falta de campos de batalla convencionales, se estaba librando en los terrenos de la economía, la propaganda y, muy especialmente, la ciencia. No es de extrañar que, al término de la conferencia, se llegara al punto de garantizar que en solo veinte años se produciría un gran avance en este campo. ¿Un aviso para navegantes?

Durante el encuentro, McCarthy acuñó el término «inteligencia artificial», y Minksy dejó constancia de su convencimiento de que los ordenadores llegarían a superar de forma extraordinaria a la inteligencia humana. Desde luego, no era el primero en concebir esa posibilidad. El propio Turing escribió que no solo era probable, sino que tendría como consecuencia que las máquinas tomaran el control. El mismo año de la reunión de Dartmouth, el popular es-

critor de ficción especulativa Isaac Asimov publicaba uno de sus relatos más célebres, *La última pregunta*, en el que una IA capaz de aprender por sí misma guiaba a la humanidad a medida que se expandía por el universo.

Por un lado estaba un enfoque del estilo *top-down* (o «de arriba abajo»), liderado por Simon y el informático Allen Newell, que consideraba la inteligencia como una propiedad independiente del soporte, biológico o de otro tipo, y que además podía describirse en términos simbólicos o abstractos. Una vez adecuadamente descrita en dichos términos, una máquina diseñada específicamente para ello podía llevar a cabo el procesado simbólico y pasar a ser inteligente. Este enfoque, que priorizaba el modelado abstracto de la inteligencia (el «arriba»), opinaba que había que olvidarse del cerebro y centrarse en los sistemas simbólicos y las reglas para manipularlos. A dicho enfoque se oponía otro del tipo *bottom-up* (del inglés «de abajo arriba»), liderado por el psicólogo Frank Rosenblatt, que defendía que la estructura específica del cerebro era un elemento fundamental para la emergencia de la cognición y, por tanto, el primer paso era diseñar un *hardware* que imitara dicha estructura (el apelativo «de abajo arriba» obedece a la insistencia de partir de lo concreto a lo abstracto). La sana competencia entre ambas escuelas de pensamiento, la simbólica y la conexionista, respectivamente, caracterizaría la primera Edad de Oro de la inteligencia artificial, un breve periodo que se extendió hasta mediados de la década de 1960.

El constructor de teoremas

Tal vez el fruto temprano más espectacular del enfoque simbólico fue la creación, en 1955, de Logic Theorist (del inglés «el teórico lógico»), un programa informático desarrollado por Newell, Simon y el programador Cliff Shaw, que podía, por primera vez, demostrar

teoremas matemáticos. Por haber sido escrito con la intención de simular la capacidad humana de resolver problemas, está considerado el primer programa de inteligencia artificial de la historia. Con el tiempo, Logic Theorist acabó demostrando 38 de los primeros 52 teoremas postulados en los célebres *Principia Mathematica* (1910-1913) de Bertrand Russell y Alfred North Whitehead.

Logic Theorist se valía de axiomas y de unas reglas de derivación para, aplicando su fuerza computacional, llegar a conclusiones que eran, por definición, lógicamente válidas. Veamos un ejemplo sencillo con dos premisas y una única regla de derivación:

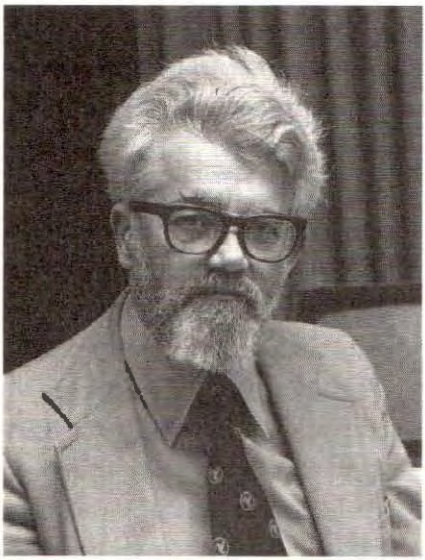
Premisa 1: Sócrates es un hombre.
Premisa 2: Los hombres son mortales.
Regla de derivación: Si ambas premisas son ciertas, entonces lo es también que «Sócrates es mortal».

Para que Logic Theorist fuera capaz de realizar manipulaciones como estas, primero hay que formalizar los términos de forma simbólica, por ejemplo:

A: Sócrates
B: hombre
$A \rightarrow B$ (= «Socrates es un hombre»).
C: mortal
$B \rightarrow C$ (= «Los hombres son mortales»).

Ahora, introducimos la regla de derivación, que en este caso es una forma lógica conocida como *argumento cadena*, y que afirma que:

Si $A \rightarrow B$ y $B \rightarrow C$, entonces $A \rightarrow C$.
Y ya hemos obtenido nuestro teorema, que no es otro que «Sócrates es mortal».



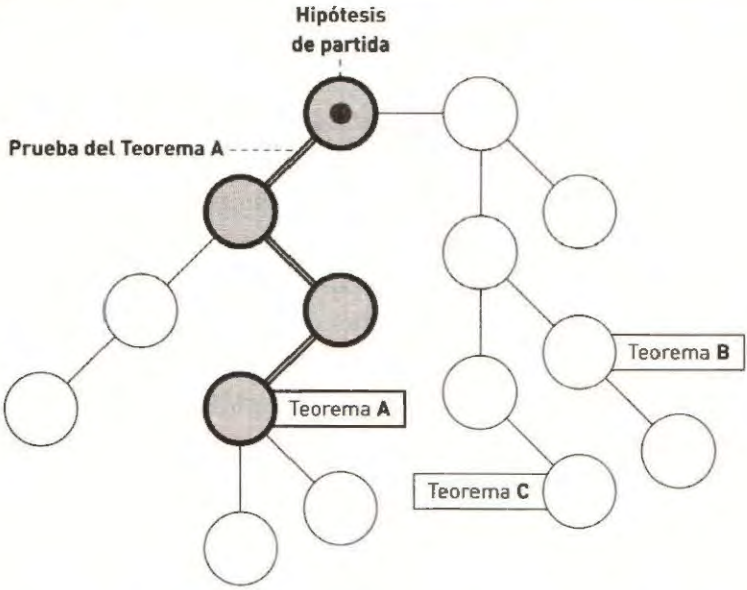
— Arriba a la izquierda, John McCarthy en la Universidad de Stanford, a la que se incorporó en 1962. A su derecha, Frank Rosenblatt, el padre del perceptrón, en Cornell a finales de los años sesenta. Abajo, dos de los creadores —junto con Cliff Shaw— de Logic Theorist, Herbert A. Simon (izq.) y Allen Newell.

Logic Theorist se valía de múltiples reglas de derivación que, aplicadas a las premisas que sus programadores habían introducido, iban generando nuevas premisas lógicamente válidas (podemos imaginar las distintas premisas como ramificaciones de un árbol cada vez más complejo). En paralelo a la generación de premisas, el programa las iba examinando para reconocer en ellas un teorema de entre los comprendidos en los *Principia*. Llegado el caso, daba una señal: ¡Eureka! Este enfoque, sin embargo, tiene un problema importante. A medida que se multiplican las premisas de partida y las reglas de derivación, los árboles de potenciales teoremas alcanzan una complejidad creciente y muy pronto intratable, como puede intuirse al observar la figura 2. ¡Demasiadas premisas entre las que buscar! Para solucionar este problema, Newell, Simon y Shaw introdujeron atajos, prohibiciones y otras técnicas *ad hoc* que impedían a Logic Theorist generar «ramas» poco prometedoras. Cada una de estas técnicas se denomina *heurística*, del griego «hallar» o «descubrir», un término que hizo fortuna.

El inconveniente de las técnicas heurísticas es que se basan en valoraciones previas sobre lo que puede ser un camino poco prometedor en el contexto de la tarea que se debe realizar. Y esas valoraciones son, en parte, intuitivas. A pesar de esto, o precisamente por esto, condicionan en gran medida la calidad del sistema inteligente al que se aplican.

Logical Theorist fue, en cualquier caso, un logro extraordinario que dio alas al paradigma simbólico en estos años. McCarthy y Minsky, por ejemplo, fundaron el Laboratorio de Inteligencia Artificial del MIT, en 1959, con la intención de investigar en esta vía, lo que dio un gran impulso a la disciplina. Sin embargo, este optimismo empezaba a encontrar también su contrapeso en algunas voces que señalaban las dificultades del camino hacia la IA. Una de ellas era la de John von Neumann, que en su opúsculo *El ordenador y el cerebro*, publicado póstumo en 1958, señalaba que, a

Fig. 2



Esquema de un árbol de teoremas a partir de una hipótesis de partida como los generados por Logic Theorist. La suma de nodos que conduce a un teorema cualquiera constituye la demostración de dicho teorema.

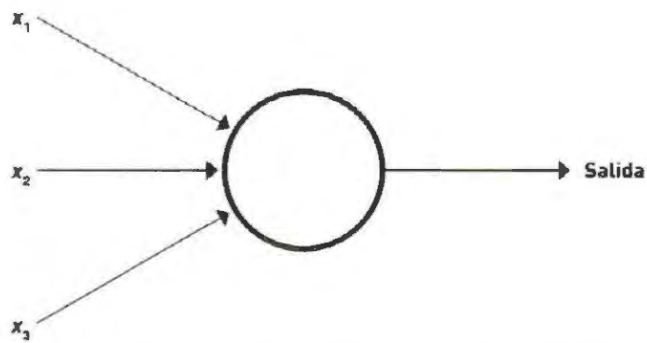
pesar de las similitudes que guardaban el cerebro humano y el ordenador, estos diferían en aspectos esenciales, como la potencia de computación o la capacidad del cerebro para procesar diversas tareas en paralelo. Sería un primer toque de atención contra el optimismo de la Edad de Oro.

Neuronas artificiales

Los partidarios del enfoque *bottom-up*, o conexionistas, no se quedaron de manos cruzadas, y pronto desarrollaron conceptos tan importantes como los de neurona artificial y red neuronal.

Los fundamentos teóricos, tanto de las neuronas artificiales como de las redes neuronales, fueron establecidos en 1957 por Rosenblatt, que desarrolló sus tareas de investigación en la estadounidense Universidad de Cornell apoyándose en los trabajos previos de Warren McCullough y Walter Pitts. El objetivo de Rosenblatt era aproximarse a la acción cognitiva del cerebro humano mediante copias artificiales de las neuronas, a las que bautizó como *perceptrones*. Dichas neuronas artificiales podrían conectarse entre sí y formar redes neuronales de modo similar a como las neuronas biológicas componen el tejido cerebral. Estas primeras redes son el germen de las complejas estructuras que hemos examinado en el primer capítulo, y que revisitaremos más adelante.

El perceptrón de Rosenblatt es un programa que tiene como entrada un vector binario x_1, x_2, x_3... (es decir, una ristra de longitud variable de ceros y unos), a partir de la cual produce una única salida también binaria:



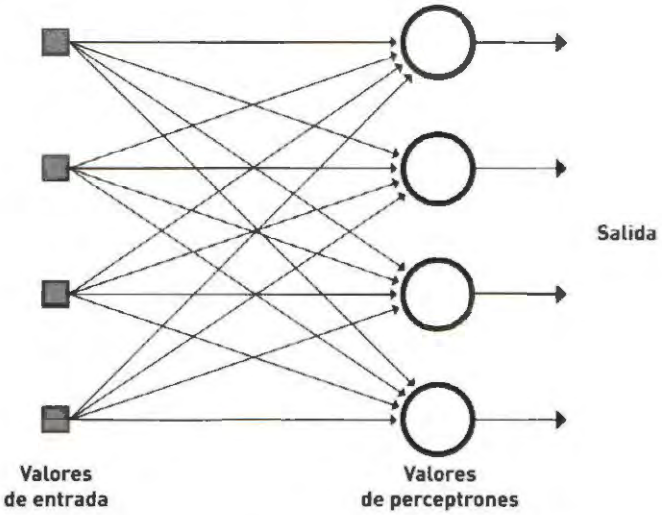
Para calcular cuál será esa salida, también llamada *resultado del perceptrón*, cada señal de entrada (x_1, x_2 y x_3, en el ejemplo) es ponderada con un peso (w_1, w_2 y w_3) cuyo valor, mayor o menor, refleja la importancia de cada señal de entrada. El resultado del perceptrón es 1 si la suma ponderada de las señales de entrada con sus pesos es mayor que un límite determinado o *umbral* y 0 si es igual o menor:

$$\text{Salida} = \begin{cases} 0 & \text{si } \sum_j w_j x_j \leq \text{Umbral} \\ 1 & \text{si } \sum_j w_j x_j > \text{Umbral} \end{cases}$$

Cuando el valor de salida es 1 se dice, como en el caso de las neuronas biológicas, que la señal se propaga.

En definitiva, una neurona artificial recibe uno o varios valores de entrada y, en función del peso asignado a cada uno de esos valores, genera o no un valor de salida. Las semejanzas entre las neuronas artificiales y sus hermanas biológicas son evidentes: las entradas harían las veces de las dendritas, el cálculo de la salida se haría en el equivalente al soma, y la señal se propagaría por una ramificación de salida única similar al axón.

Es posible combinar varias neuronas artificiales en forma de red para que computen conjuntamente determinados valores de entrada:



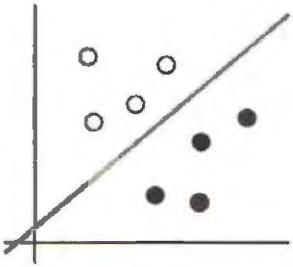
La aportación genial de Rosenblatt fue introducir la variable de los pesos. Esto permitía ir ajustando los valores de salida de los perceptrones para que fueran acercándose cada vez más al resultado buscado. Es por esta capacidad de ir mejorando progresivamente

el rendimiento en función de los resultados obtenidos por lo que se dice que la red neuronal aprende. Más adelante, se diseñaron formas de que dicho ajuste se realizara automáticamente, lo que aumentó muchísimo la capacidad y velocidad de aprendizaje de las redes neuronales, como veremos en el capítulo 4.

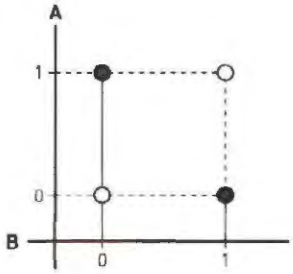
Si Turing hubiera estado vivo por aquellos años, caben pocas dudas de que sus simpatías habrían estado del lado conexionista, dado el énfasis de este enfoque en el aprendizaje como vía fundamental para alcanzar el comportamiento inteligente (recordemos su propuesta de una máquina infantil). Las redes neuronales permiten, además, alterar las reglas de aprendizaje en función de si sus resultados son mejores o peores, lo que a su vez evoca esa suerte de evolución dirigida que el británico consideraba tan prometedora.

¿Qué tipo de problemas son capaces de resolver las redes neuronales? Fundamentalmente, los de reconocimiento de patrones en un conjunto de datos. Y si de algo hay en abundancia gracias a la explosión de internet son, precisamente, datos. Son tan buenas en ello que en la actualidad tienen un papel protagonista en ámbitos tan diversos como la identificación de imágenes o sonidos, el control de vehículos, la diagnosis médica o la toma de decisiones en entornos reglados. Tal es su potencial en toda clase de ámbitos que ya se habla de un nuevo paradigma de inteligencia artificial basado en los datos.

Sin embargo, el camino hacia la satisfacción de ese enorme potencial no fue fácil. En 1969, Minsky y Seymour Papert publicaron un artículo, «Perceptrones», en el que señalaban una, a sus ojos, importantísima carencia de los perceptrones y, por consiguiente, de las redes neuronales construidas basándose en ellos. Supongamos que existen dos tipos de datos, blancos y negros. Si la red neuronal funciona de forma correcta, será capaz de identificarlos adecuadamente y, por tanto, separarlos como hace en esta gráfica la línea recta:



Supongamos ahora unos datos singulares cuya representación gráfica sea la siguiente:



En este caso, no existe recta alguna que podamos dibujar que separe netamente unos tipos de datos de otros. Se dice entonces que los datos no son separables linealmente. Pues bien, resulta que las primitivas redes neuronales, formadas por una sola capa de perceptrones, no eran capaces, como demostraron Minsky y Papert, de resolver este problema.

¿Por qué supone eso un grave inconveniente? La figura de más arriba es la expresión gráfica de una función llamada XOR. Dados unos valores de entrada A y B, la función XOR da un valor de salida 0 o 1, según la tabla adjunta:

Entrada A	Entrada B	Salida
0	0	0
0	1	1
1	1	1
1	0	0

Esta función, al estar en lenguaje binario, tiene muchos usos en programación. Así, otro modo de exponer el problema detectado por Minsky y Papert es que las redes neuronales de una sola capa no eran capaces de entender programas que exigieran computar una operación XOR. El llamado problema XOR tuvo en jaque a los desarrolladores de redes neuronales durante un tiempo.

Entretanto, a la escuela simbólica las cosas no le iban mucho mejor. Durante más de una década, el Gobierno estadounidense había invertido en el desarrollo de programas de traducción automática, dadas las ventajas estratégicas de disponer de un modo rápido y fiable de grandes cantidades de documentos traducidos del ruso al inglés (por razones bélicas obvias). Pero, hacia 1966, comprobaron que las expectativas estaban lejos de cumplirse. Los investigadores habían advertido que reducir el modo en el que el lenguaje genera significado a un número manejable de reglas resultaba casi imposible, ya que estas no eran capaces de recoger la información de fondo y de contexto que manejan los hablantes y que es esencial para que se entiendan. Entre la información de fondo más obvia están las metáforas, las frases hechas o los sobreentendidos. Sin una adecuada comprensión de su naturaleza y funcionamiento, todos los intentos de la escuela simbólica por capturar el sentido de las frases con mecanismos como los empleados por Logic Theorist estaban condenados al fracaso. Entre las limitaciones de unos y de otros, el campo de la inteligencia artificial dejó de atraer fondos, adentrándose en lo que se conoció como «el invierno de la IA». Dicho estancamiento duraría hasta mediados de la década de 1980.

SISTEMAS EXPERTOS Y REDES NEURONALES MULTICAPA

Los problemas de la IA no se limitaban al problema XOR y al de la información de fondo. Había una creciente desesperanza entre

> MARVIN MINSKY, EL CEREBRO DE LA IA

Si el honor de padre de la IA no estuviera reservado a Alan Turing, Marvin Minsky hubiera sido un excelente candidato. Nacido en Nueva York en 1927, Minsky se doctoró en matemáticas en la Universidad de Princeton. En 1958 ingresó en el Massachusetts Institute of Technology, en cuya facultad permaneció hasta su muerte en 2016. Todas sus aportaciones al campo de la IA, desde la teoría de marcos hasta sus influyentes críticas a las redes neuronales, delatan su convencimiento de que el cerebro humano era solo un tipo especialmente complejo de máquina. En su obra seminal *La sociedad de la mente* (1986), Minsky postuló que la inteligencia era fruto de la acción combinada de agentes semiautónomos ellos mismos carentes de inteligencia. Además de sus aportaciones teóricas destacan invenciones como el primer casco de realidad virtual, por el que obtuvo una patente, y el microscopio confocal. En paralelo a su actividad científica, ejerció un papel extraordinario como dinamizador de la investigación en computación y fue uno de los fundadores del MIT Media Lab.

— Marvin Minsky, en la década de 1980, con un guante robótico interactivo.

los partidarios del enfoque simbólico por las desmesuradas exigencias en términos de capacidad de cálculo de cualquier modelo que quisiera capturar un proceso mental mínimamente complejo.

> De tres a ocho años tendremos una máquina con una inteligencia similar a la de un ser humano medio.
>
> MARVIN MINSKY (1970)

Se calculó, por ejemplo, que un ordenador que fuera capaz de reproducir algunas de las funciones de la retina humana requeriría poder computar unos mil millones de operaciones por segundo. Sin embargo, el Cray-1, el ordenador más potente de la época, tenía una capacidad de 160 millones de operaciones por segundo. Para hacernos una idea de las limitaciones tecnológicas a las que los pioneros de la IA tenían que enfrentarse, un *smartphone* de gama media alcanza, a finales de la segunda década del siglo XXI, una potencia superior a los 100 000 millones de operaciones por segundo. ¿Y qué decir de las exigencias combinatorias de juegos como el ajedrez, que puede presentar alrededor de 10^{50} posiciones posibles? Pero no solo era una cuestión de potencia de cálculo. El enfoque simbólico se topaba una y otra vez con su incapacidad para capturar las acciones cognitivas más elementales del ser humano. Supongamos el caso de un individuo que ve a un conocido por la calle. Esta operación implica diferenciar una imagen de otras del entorno, identificarla como un rostro y asignar ese rostro a una persona concreta. Para este conjunto de acciones, un ser humano se toma una fracción de segundo. Pero el programa necesario tardaría miles de años en ejecutarse.

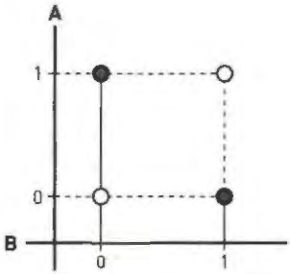
Y no solo hablamos de las acciones cognitivas del ser humano. Un insecto, como una simple mosca, es capaz de orientarse y desplazarse en el espacio computando, en apariencia de forma inmediata, millones de datos relativos a la velocidad del aire, la forma y cercanía de los objetos situados en su campo visual, entre otros. Ambas son formas de un mismo problema, conocido como

paradoja de Moravec en honor a Hans Moravec, pionero de la IA. La paradoja se da al constatar que modelar acciones en apariencia sencillas pero que implican la intervención de los sentidos o del aparato motor, como reconocer un rostro o desplazarse por el entorno, requieren mucho más poder de computación que la simulación de ciertos razonamientos abstractos. En palabras del propio Moravec, «es comparativamente fácil conseguir que una computadora emule a una persona adulta a la hora de responder un test de inteligencia o jugar a las damas, pero resulta casi imposible dotar a la máquina de las habilidades motoras y perceptivas de un niño de un año».

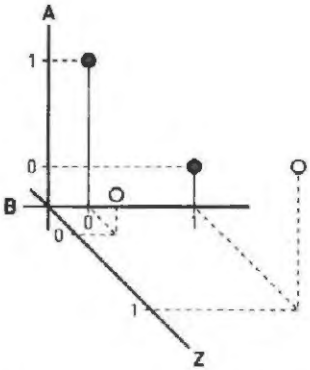
Tanto la escuela simbólica como la conexionista lograron escapar parcialmente del marasmo mediante el desarrollo de distintas tecnologías. La primera, mediante la creación de los sistemas expertos. Y la segunda, a través de la mejora de las redes neuronales hacia redes neuronales multicapa.

Los sistemas expertos, que se explorarán en detalle más adelante, son una evolución de los programas de búsqueda circunscritos a ámbitos específicos, donde el rango de decisiones posibles es más estrecho. Pronto demostraron su utilidad en ámbitos tan diversos como el diagnóstico médico o la predicción de factores de riesgo en poblaciones.

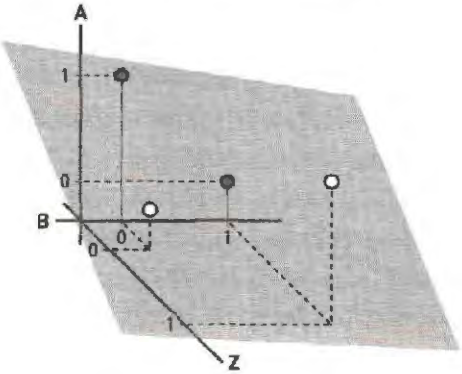
Por su parte, la escuela conexionista se anotó un tanto importante gracias al diseño de las redes neuronales multicapa, es decir, formadas por varias capas de perceptrones. Este diseño, unido al algoritmo de *backpropagation* o retropropagación, que dotaba a las redes neuronales de la capacidad de ajustarse de forma automática, dio por fin carpetazo al problema XOR. Las capas adicionales proporcionaban a la red neuronal la capacidad de manejar información en más dimensiones y la retropropagación permitía que una red neuronal así diseñada aprendiera de forma más rápida y eficiente. Recordemos la expresión gráfica de la forma lógica XOR:



Cuando la red neuronal es del tipo multicapa es capaz de añadir una dimensión adicional al problema, de manera que se genera un tercer eje Z tal que:



Ahora, ya es posible separar unos datos de otros; no con una recta bidimensional, pero sí con un plano tridimensional (el plano quedaría entre el lector y los dos puntos negros):



Con este nuevo diseño, las redes neuronales daban un paso de gigante hacia una creciente complejidad y potencia.

CUANDO LA IA GANÓ EN LOS JUEGOS DE INTELIGENCIA

Cuando ya se apagaba el siglo XX, como si de un heraldo del futuro se tratara, la IA atrajo de nuevo la atención de los medios al superar, por primera vez, al ser humano en una actividad cognitiva compleja: un juego de inteligencia. En 1996, Deep Blue, un ordenador desarrollado por IBM, derrotaba al campeón mundial de ajedrez Garri Kaspárov. A pesar de lo espectacular del resultado, el avance que supuso para la IA no fue tan importante como cabría esperar. En realidad, Deep Blue no había vencido a Kaspárov porque fuera verdaderamente inteligente: lo hizo básicamente gracias a la fuerza bruta computacional, pues podía evaluar 200 millones de posiciones por segundo y compararlas con 700 000 jugadas anteriores de grandes maestros. Kaspárov, de hecho, desdeñó aquella victoria aduciendo que Deep Blue solo era inteligente como podía serlo un reloj despertador programable. Y, sin embargo, lo cierto es que algunas capacidades que creíamos exclusivas de la inteligencia son simulables por un ordenador que, además, al desplegarlas acaba siendo más competente que ningún humano. Esta realidad se ha puesto de manifiesto una y otra vez en juegos como el Scrabble, las damas o los crucigramas. Tal vez, como sugiere Bostrom, se trate de tímidos avances que podrían rápidamente ser superados gracias al descubrimiento de algún tipo de algoritmo muy sencillo:

> La maestría en ajedrez resultó ser alcanzable por medio de un algoritmo sorprendentemente simple. Resulta tentador especular que otras capacidades (tales como la capacidad de razonamiento general o alguna habilidad clave implicada en programación) podrían

ser igualmente alcanzables a través de algún algoritmo sorprendentemente simple.

Lo que Deep Blue no significó en términos de un auténtico avance, sí lo hizo AlphaGo, una IA concebida para jugar al go, probablemente el juego de mesa más complejo jamás diseñado. Baste decir que en su tablero hay más posiciones posibles que átomos en el universo. Inteligencia, estrategia, creatividad e intuición son algunas de las capacidades que ha de tener todo jugador de go. Pues bien, AlphaGo consiguió en 2016 derrotar en cuatro de cinco partidas al considerado mejor jugador del mundo, el surcoreano Lee Sedol. A diferencia de Deep Blue, AlphaGo sacó partido de las más avanzadas técnicas de IA y en particular del *deep learning*, una evolución de las redes neuronales que exploraremos en el capítulo siguiente. Gracias a esta tecnología, que exige una capacidad de cálculo solo recientemente disponible, AlphaGo aprendió de manera autónoma partiendo de la información acumulada en cientos de miles de partidas registradas y pasando dos años jugando contra ella misma una y otra vez.

¿Consiguieron Deep Blue o AlphaGo superar el test de Turing? La respuesta es que, posiblemente, no. Aunque Turing nunca lo especificó así, el supuesto implícito es que el ordenador debe simular a un humano en situaciones más generales que las que plantea un torneo de ajedrez o de go. Sin embargo, quizá no podría decirse lo mismo en el caso de Watson, el sistema informático de IBM que, en 2011, derrotó a los campeones mundiales de Jeopardy! El concurso televisivo planteaba un juego en el que los participantes tenían que deducir, a partir de una pista, una respuesta que, a su vez, debía darse en forma de pregunta. Estas pistas a menudo contenían dobles sentidos o juegos de palabras, como por ejemplo esta: «Con mucha gravedad, este becario del Trinity College se convirtió en profesor de la Cátedra Lucasiana de Matemáticas de la Universidad de Cambridge en 1669». En este caso, Watson

debía responder «Isaac Newton». El lector habrá reconocido en la palabra *gravedad* un ejemplo del problema de la información de fondo mencionado anteriormente, y también habrá tomado nota de que Watson fue capaz de resolverlo. Pero ¿no era esta una capacidad solo reservada a los seres humanos?

> El mundo es un problema de datos.
>
> ANDREW MCAFFEE

Aunque estaba conectado a internet mientras jugaba, Watson tenía igualmente memorizadas 200 millones de páginas. Sin embargo, la potencia de Watson no reside en su memoria, y ni siquiera en su capacidad de cálculo, sino en ser capaz de buscar en el seno de tanta información de forma eficiente y saber cuándo ha encontrado la respuesta correcta. Además, por supuesto, de ser capaz de entender el lenguaje natural en el que se formulaban las preguntas y de comunicarlas de forma inteligible a la audiencia.

Cuando finalmente el campeón de Jeopardy! a lo largo de setenta y cuatro programas consecutivos Ken Jennings fue derrotado por Watson, asumió su derrota con deportividad con estas palabras: «Por mi parte, doy la bienvenida a nuestros nuevos amos, los ordenadores».

Parece claro que estamos no solo dejando atrás el invierno de la IA, sino plenamente inmersos en su primavera. Sorprendentemente, este florecimiento no viene dado por ningún descubrimiento revolucionario, sino que es el resultado de una serie de factores que permiten que los algoritmos que se desarrollaron hace décadas se puedan utilizar con todo su potencial en aplicaciones concretas. Estos factores son básicamente tres:

1. El *big data*: la proliferación de los datos disponibles con el auge de internet ha creado un inmenso repositorio de miles de millones de documentos, vídeos e imágenes que permiten entrenar redes neuronales y sus derivados a un nivel inédito hasta el momento.

2. La computación en la nube o *cloud computing*: esta tecnología ha aumentado la capacidad de computación y se ha revelado imprescindible para entrenar a las redes neuronales. Hay que decir que si la computación cuántica llega algún día, se superará largamente la capacidad del *cloud computing*.
3. El Internet de las cosas o *Internet of Things*: la explosión de dispositivos y sensores que se encuentran conectados entre sí a través de internet no solo constituye una fuente de información por sí misma, sino que, al abrir un canal de comunicación en tiempo real con el usuario, ha disparado la demanda de aplicaciones de IA capaces de interaccionar con él.

Según la descripción hecha por IBM, Watson combina adelantos en el procesamiento de lenguajes naturales, con la representación del conocimiento, el razonamiento y el aprendizaje aplicado al campo abierto de búsquedas de respuestas. Es decir, que en él confluyen técnicas tanto simbólicas como conexionistas. Si a esta versatilidad se suman los tres factores descritos, ¿quién se atreve a poner límites a los logros futuros de la inteligencia artificial?

INTELIGENCIAS FUERTES Y DÉBILES

Hasta ahora hemos explorado las ideas fundamentales de la IA en su contexto histórico. El objetivo era mostrar su evolución, por lo que se han expuesto de forma sencilla. De aquí en adelante, sin embargo, vamos a aumentar la profundidad del análisis. Para ello nos resultará muy útil dejar clara la distinción entre inteligencia artificial fuerte e inteligencia artificial débil.

El filósofo estadounidense John Searle identificó en 1980 dos hipótesis diferentes sobre la inteligencia artificial. La primera de ellas, que caracterizó como *fuerte*, sostiene que un sistema de inteligencia artificial puede tener mente (es decir, tener consciencia

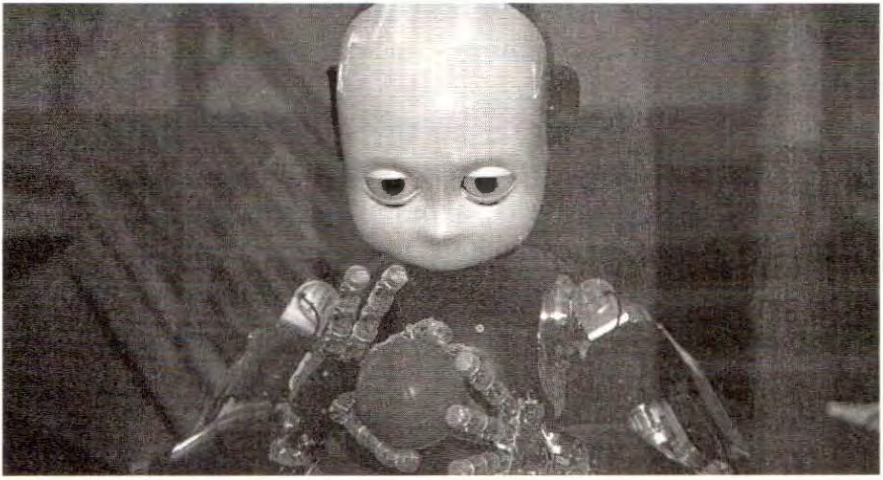
> TESTS DE INTELIGENCIA PARA MÁQUINAS

Los investigadores en IA, aficionados como son a las heurísticas, han diseñado distintas pruebas que, como la de Turing, sirven de atajo para evaluar si un sistema es inteligente:

— La prueba del café, propuesta por Ben Goertzel, científico pionero en el estudio de la IA general, mide la capacidad de la máquina para entrar en una casa y averiguar cómo hacer café. La máquina debe poder localizar la cafetera, encontrar el café, poner agua en la máquina, encontrar una taza y hacer el café.

— La prueba del estudiante universitario, también de Goertzel, valora el desempeño de un sistema que tiene que asistir a un curso universitario, aprender las lecciones y ser capaz de superar los mismos exámenes a los que se enfrentan los estudiantes.

— La prueba del empleo, propuesta de Nils Nilsson, evalúa si un sistema es capaz de realizar un trabajo económicamente importante con la misma eficiencia y cumplimiento que un humano.



— El niño robot iCub, desarrollado por el Instituto Italiano de Tecnología, es capaz de manipular objetos, como lo haría un robot que pasara la prueba del café.

e intencionalidad). La segunda, a la que bautizó como *débil*, acepta que un sistema inteligente solo puede actuar *como si* pensara y tuviera mente. Searle quería así separar entre aquellos que creían que una máquina inteligente podía desarrollar una conciencia, y aquellos que opinaban que, por muy inteligente que pudiera llegar a ser su comportamiento, una máquina no hace más que ejecutar reglas ciegamente. Ambos términos, *fuerte* y *débil*, tuvieron aceptación entre los investigadores y en la actualidad se manejan de forma generalizada, aunque con un sentido algo distinto. A quienes tienen como objetivo la construcción de una inteligencia artificial capaz de emular en todo al ser humano se les adscribe el deseo de construir una IA fuerte, mientras que aquellos que se contentan con crear sistemas capaces de mostrar inteligencia a la hora de solucionar tareas específicas trabajan bajo el paradigma de la IA débil. Démonos cuenta de que la distinción no entra a considerar si la máquina tiene o no mente, es decir, se ha perdido la que era la intención principal de Searle a la hora de introducir el término.

Como se ha podido ver, la IA es una disciplina muy joven, con poco más de seis décadas de historia. Por ello, su definición y clasificación carecen de la robustez de otras disciplinas científicas con más recorrido. En otras palabras, aún no hay un paradigma o teoría unificada de la IA, lo que favorece constantes vaivenes en su ámbito de estudio. Por ejemplo, lo que hace solo unas décadas era IA, como el razonamiento automático del tipo del Logic Theorist o el de algunos sistemas expertos, ahora ya se considera computación tradicional. Esta falta de concreción tiene como causa subyacente que tampoco la inteligencia, la que se supone es el verdadero objetivo de la disciplina, está adecuadamente definida. Ahora bien, como hemos visto en el caso de la distinción entre fuerte y débil, esta ausencia de precisión en el corazón mismo de la inteligencia artificial no tiene por qué ser un obstáculo insalvable. La mayoría de los investigadores en IA están de acuerdo en las propiedades mínimas que todo sistema inteligente fuerte debe cumplir, y eso

les basta. Estas propiedades están directamente inspiradas en las habilidades de las inteligencias biológicas, y son:

1. La capacidad de tomar decisiones y planificar a medio y largo plazo.
2. La capacidad de interactuar con el mundo real percibiendo, entendiendo y actuando en consecuencia. Desde luego, orientándose, desplazándose en el espacio y manipulando objetos de acuerdo con sus fines, pero también, y en particular, relacionándose con una parte muy significativa del mundo real: nosotros. Es decir, deben reconocer y entender el lenguaje natural y relacionarse emocionalmente.
3. La capacidad de adaptarse al entorno y acumular conocimiento nuevo con el objetivo último de solucionar problemas también nuevos.

En los dos capítulos siguientes vamos a explorar las distintas estrategias, ya se incluyan en el paradigma fuerte o en el débil, que han de llevar a las IA del futuro a exhibir estas capacidades. Es decir, las orientadas a conseguir máquinas que razonen, se comuniquen y aprendan.

03 MÁQUINAS QUE RAZONAN E INTERACTÚAN

Emular la razón humana ha sido uno de los objetivos históricos de la IA. Pero gana enteros la noción de que la inteligencia necesita, además, de un entorno con el que interactuar.

Es muy probable que el lector, al imaginar una IA, piense en un gran ordenador inmóvil conectado al exterior mediante infinidad de cables y antenas y que no emite más ruido que el de su maquinaria interna. Y, sin embargo, muchos expertos opinan que la inteligencia es inseparable del cuerpo que habita (o, al menos, de la experiencia de habitar un cuerpo). En palabras del científico cognitivo británico Andy Clark,

> Los cerebros biológicos son, principalmente, centros de control de los cuerpos biológicos. Y los cuerpos biológicos se mueven y actúan en el mundo real.

En la parte que sigue vamos a liberar a nuestra IA de su aislamiento y le vamos a proporcionar visión, habla e incluso movimiento. Es decir, vamos a dotarla de autonomía para desenvolverse en un entorno. El término riguroso para una IA de estas características es *agente*. Cuando un agente es capaz de actuar para satisfacer unos fines se dice de él que es *racional*. Un agente racional capaz de lle-

var a cabo tareas complejas en el mundo real, tales como graduarse en una universidad o desempeñar un trabajo humano, satisfaría un criterio de inteligencia general todavía más exigente que el de Turing.

> Hay que tomar como modelo el comportamiento humano, no un ideal de pensamiento.
>
> PETER NORVIG

Hay distintos tipos de agente en función de varios aspectos. En primer lugar, se distingue entre *agentes reactivos simples* o *agentes reactivos basados en modelo*. Los primeros responden a sus percepciones, sin más; los segundos, en cambio, se sirven de sus percepciones para actualizar con ellas un modelo del entorno, en función del cual deciden qué acciones tomar. Otra distinción clave es la que atañe a la motivación de los agentes. Los *basados en objetivos* actúan para alcanzar sus objetivos; los *basados en la utilidad* actúan para maximizar la utilidad esperada, donde por utilidad podemos entender felicidad o bienestar. Los agentes reactivos basados en modelo y motivados por la maximización de la utilidad serían aquellos que de forma más cercana estarían emulando el comportamiento humano. Tanto unos como otros son susceptibles de mejorar su desempeño mediante el aprendizaje, un aspecto fundamental del que nos ocuparemos en el capítulo siguiente.

De la definición de agente (excepto la del meramente reactivo), se desprende que su inteligencia ha de permitirle, al menos, conocer su entorno; actualizar su conocimiento en función de nueva información; planificar una acción; y decidir qué acción tomar. Llamaremos a la suma de todas estas capacidades *razonar*. No se dispone en la actualidad de un único modelo ampliamente aceptado acerca de cómo razona el cerebro humano (la elusiva «arquitectura cognitiva» que citaban los expertos en su encuesta), así que los investigadores en IA se han tenido que contentar con diseñar estrategias que intenten emular el raciocinio en ámbitos específicos. La que vamos a ver es, por tanto, IA aplicada o débil, basada en su caso en el análisis formal y estadístico del comportamiento humano.

LA RAZÓN COMO BÚSQUEDA DE SOLUCIONES A UN PROBLEMA

La única exigencia implícita en la noción de agente racional en lo que concierne a la inteligencia es que esta debe estar orientada a la acción, es decir, debe ser una inteligencia eminentemente práctica. Es por ello por lo que buena parte de los esfuerzos de los investigadores se han orientado a diseñar sistemas que apliquen la razón a la resolución de problemas.

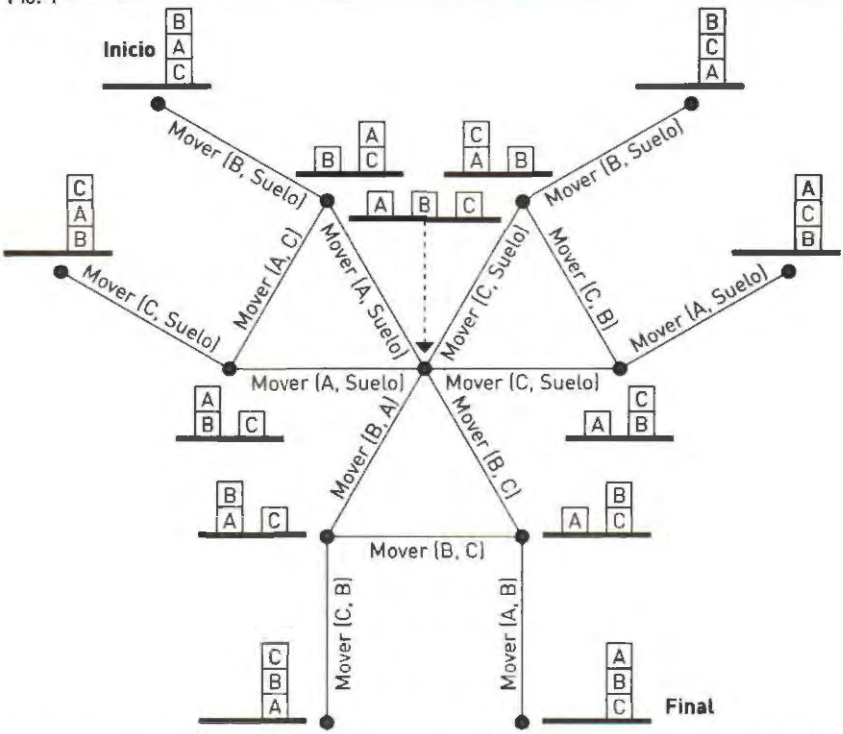
El modelo formal que subyace a estos sistemas es el de búsqueda, en el sentido de que hallar la solución consistirá en buscar la secuencia de acciones que conduzcan al objetivo deseado. Expresado de forma todavía más abstracta, la búsqueda es el proceso de descubrir y agrupar aquella secuencia (o secuencias) de acciones que llevan de un estado cualquiera a ese estado que representa el objetivo perseguido. Una solución es, en este planteamiento, una secuencia de acciones que permiten llegar de un estado inicial a otro final. Un problema de búsqueda en IA, en resumen, consta de: un conjunto o *espacio de estados*, un conjunto de *acciones posibles* u *operadores*, un punto de partida de la búsqueda o *estado inicial* y un medio para comprobar que un estado cualquiera es, en efecto, solución al problema, o *función objetivo*.

Para analizar los problemas de búsqueda se emplea una forma matemática de representar la información conocida como *grafo implícito* o *grafo de espacio de estados*. Un grafo de este tipo se compone de *nodos* y *arcos*, que representan los estados y los operadores, respectivamente (fig. 1).

Otra forma posible de representarlo es mediante un *grafo explícito* o *árbol de búsqueda*, que no es otra cosa que el subgrafo del grafo implícito que se genera, paso a paso, durante el proceso de búsqueda (fig. 2).

El grafo implícito representa el problema que hay que resolver; y el árbol de búsqueda, todas las posibles soluciones al problema. Por

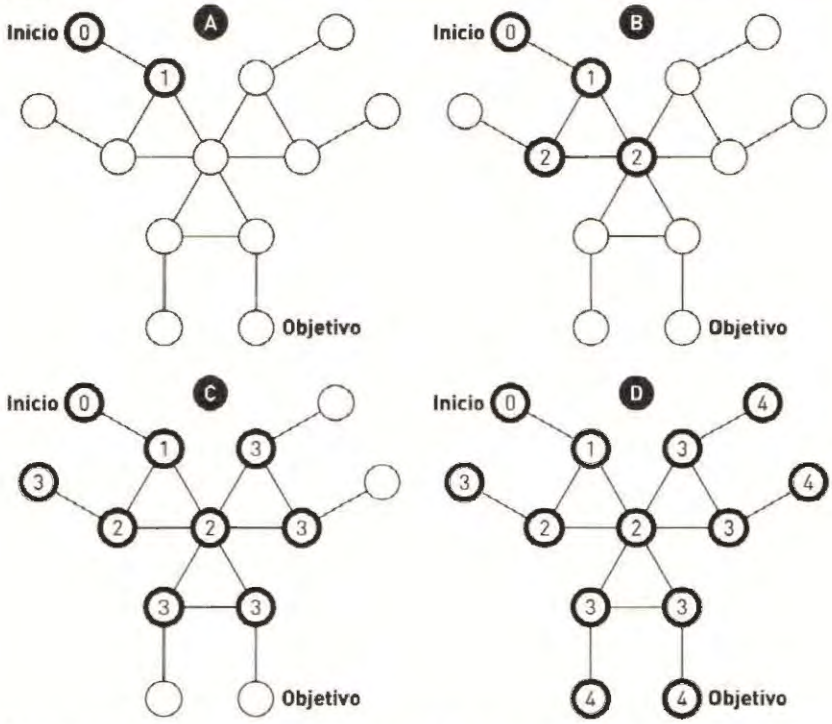
Fig. 1



Supongamos un juego consistente en disponer en columna tres bloques A, B y C moviendo solo un bloque por turno. Este grafo de estados recoge todos los posibles movimientos tomando como posición de partida la columna BAC del extremo superior izquierdo y como objetivo la columna ABC del extremo inferior derecho.

tanto, el camino de un estado inicial cualquiera al estado final en un árbol de búsqueda es, en sí, la solución al problema. Ahora bien, generar todo el árbol de búsqueda para encontrar un estado final es un problema intratable desde el punto de vista de la complejidad computacional que conlleva. El término «intratable» se emplea aquí en su acepción técnica: un problema es intratable cuando su tamaño aumenta linealmente pero el tiempo de ejecución del algoritmo que lo expresa lo hace exponencialmente. Los problemas intratables

Fig. 2



Árbol de búsqueda o grafo implícito del grafo de estados anterior, en el que se observa que pasar de la posición de inicio a la posición objetivo requiere cuatro pasos. Los pasos concretos que conducen de una posición a la otra constituyen la solución al problema planteado originalmente.

pronto colman la capacidad de cómputo de un ordenador, incluso en el supuesto de que esta crezca al ritmo marcado por la ley de Moore.

Como ya hemos dicho, hay muchos problemas prácticos que se pueden formular como un problema de búsqueda en su árbol de estados. Son de este tipo todos los que implican hallar la ruta más eficiente entre dos puntos en una red de puntos conectados entre sí, por ejemplo, para mandar señales de vídeo entre ordenadores conectados en red o para guiar aviones en espacios aéreos densos.

También lo son los problemas combinatorios, es decir, aquellos formados por elementos que se pueden combinar entre sí, como en el juego del ajedrez.

El algoritmo más simple para encontrar la solución en un árbol de estados es la búsqueda exhaustiva o «por fuerza bruta», que consiste en explorar sistemáticamente todos los posibles estados hasta encontrar un estado final.

El ajedrez ha sido utilizado y estudiado como un tipo de problema concreto que requiere inteligencia y que puede generalizarse a otros tipos de problemas combinatorios. Aplicada a este caso concreto, la técnica de la fuerza bruta consiste en simular todas las posibles jugadas que el jugador representado por la máquina puede realizar en un momento determinado de la partida. Para cada una de estas posibles jugadas, también se analizan todas las jugadas que puede llevar a cabo el otro jugador, y así sucesivamente.

Este enfoque conlleva simular un número exponencial de jugadas. En un momento dado de una partida estándar hay unas 30 posibles jugadas diferentes. Si consideramos solo tres turnos por cada jugador, estamos abordando del orden de 10^9 (1000 millones) combinaciones posibles. Se considera que el número máximo de iteraciones que se pueden analizar es de cinco: más allá, el número de posiciones posibles hace que el análisis sea intratable en el sentido anteriormente dicho. En efecto, su dificultad aumenta de forma lineal, a medida que se analizan uno o más turnos, pero el algoritmo de búsqueda por fuerza bruta correspondiente lo hace de forma exponencial, a razón de 30 x 30 combinaciones por turno.

Si en lugar de encontrar todas las posibles soluciones a un problema nos basta con una sola, hay varias formas de explorar el árbol de búsqueda cuya eficacia depende de la naturaleza del problema. La primera estrategia es la búsqueda en amplitud. Este método rastrea todos los nodos de un mismo nivel de profundidad antes de pasar al siguiente nivel. La segunda estrategia es la búsqueda en

profundidad. En este caso, el algoritmo agota todos los hijos de un nodo cualquiera antes de regresar nuevamente arriba.

La búsqueda heurística o informada

Una estrategia de búsqueda heurística o informada es aquella que utiliza conocimiento específico del problema más allá de su definición en estados y acciones con el objetivo de hallar la solución de forma más eficiente. En este tipo de búsqueda, los nodos del árbol de búsqueda se expanden según indique una función de evaluación. La función de evaluación de un nodo está basada en el conocimiento del problema y recibe el nombre de *función de coste*. La idea subyacente es que el proceso de búsqueda más eficiente es aquel que minimiza el coste, por lo que el algoritmo priorizará siempre los nodos en los que la función de coste valga menos.

Consideremos el caso de la partida de ajedrez. Una búsqueda heurística en este contexto consistiría en una variante de la fuerza bruta que ignora las jugadas consideradas de baja calidad, es decir, aquellas que tienen un coste alto. Así, en vez de analizar sistemáticamente todas las posibilidades, el sistema descartaría jugadas como las que ponen en peligro la reina, por ejemplo.

Los algoritmos que hemos descrito hasta ahora exploran el árbol de búsqueda de forma sistemática. Esta sistematicidad permite que, al llegar a un nodo que represente un resultado válido, averiguar el camino (el conjunto de acciones) que nos han llevado a él sea trivial. En algunos problemas, sin embargo, este camino es irrelevante y solo importa el nodo objetivo. Por ejemplo, en el conocido como problema de las ocho reinas, lo único importante es encontrar una configuración de ese número de piezas de ajedrez tal que no se ataquen mutuamente, con independencia de qué movimientos se hayan empleado para llegar a ella. La forma a menudo más eficiente de resolver este tipo de problemas es la llamada

búsqueda local. Una búsqueda local empieza evaluando un estado, para a continuación explorar los nodos adyacentes a este estado en busca de un estado objetivo.

La búsqueda restringida y la planificación

Supongamos el caso de un robot que recorra las estanterías de un almacén recogiendo paquetes, como los que emplea Amazon en sus centros logísticos. En su caso, buscar de forma sistemática paquetes por recoger e irlos apilando conlleva riesgos obvios, por ejemplo, colocar paquetes muy pesados o voluminosos sobre otros muy frágiles o diminutos. Está claro que, en este caso, la búsqueda de paquetes que recoger y apilar debe someterse a unas restricciones de peso y tamaño. O pongamos el caso de un programa encargado de colorear un mapa de países. En este caso, la restricción obvia es que dos países con frontera común no pueden rellenarse de un mismo color. Muchos de los problemas a los que las IA se enfrentan en el mundo real son de satisfacción de restricciones.

De hecho, hay una relación muy estrecha entre la búsqueda con restricciones y otro ámbito de la IA, la planificación, también incluido en la categoría general de razonamiento. La diferencia fundamental entre los algoritmos de búsqueda y los de planificación es que los segundos no solo buscan una solución, sino también todos los estados intermedios que conducen del inicial al final. El conjunto formado por la solución y los pasos que llevan a ella es el plan propiamente dicho. Una vez establecida la existencia de un plan, el sistema inteligente toma las acciones pertinentes para llevarlo a cabo. Si existe más de un plan posible, el sistema debe elegir por cuál de ellos decidirse, para lo cual aplicará criterios específicos como el tiempo, el coste, etc.

Supongamos el caso de una AI a cargo de un programa europeo de trasplantes. Una vez aparece un órgano susceptible de ser trasplan-

tado, el sistema tendrá que diseñar un plan de acción que permita al paciente con mayor necesidad de un trasplante de un órgano de ese tipo recibirlo en el menor tiempo posible y en condiciones. Un sistema tal debería tener en cuenta varias restricciones: las hay obvias, como que los posibles receptores deben serlo del órgano disponible, o que el tiempo transcurrido entre la disponibilidad del órgano y el momento del trasplante debe estar por debajo de un máximo; pero también se ha de intentar minimizar los costes logísticos de la operación (el dinero de la sanidad pública no es infinito) y respetar las distintas normativas estatales y locales. Este ejemplo tiene al menos dos características importantes. La primera es que, claramente, hay dos tipos de restricciones: las inviolables, como por ejemplo la coincidencia entre el órgano y las necesidades del paciente o el límite máximo temporal entre disponibilidad y trasplante, y las que no lo son, como la minimización del coste. Decidir qué restricciones son inviolables es un aspecto clave a la hora de diseñar un sistema de planificación complejo inteligente. La segunda característica es que el problema de planificación implica tener en cuenta ámbitos distintos, por ejemplo, el del alcance geográfico de las normativas europeas, estatales y locales. Es por ello por lo que estos sistemas están estructurados en varios niveles, cada uno de los cuales está a cargo de un sistema inteligente distinto. Todos estos sistemas se coordinan de manera armónica hasta el punto de poderse suplir unos a otros. No obstante esta flexibilidad, la planificación convencional es más útil cuanto más estático es su dominio de actuación, es decir, cuando el entorno se mantiene estable a menos que el sistema planificador actúe sobre él.

LA RAZÓN COMO GESTIÓN DEL CONOCIMIENTO

A fin de que un agente se sirva de sus conocimientos para tomar una decisión adecuada, estos deben estar correctamente estructurados y ser fácilmente accesibles. Por «conocimiento» hay que en-

tender el modelo del entorno del que hablábamos antes al describir los distintos tipos de agente. Un agente de este tipo requiere de unos axiomas de partida que actúen como *base* de ese conocimiento y de un *motor de inferencias* que sea capaz de generar (inferir) conocimiento nuevo a partir de axiomas nuevos. Para que todo ello sea posible, el conocimiento tiene que estar expuesto en un lenguaje que lo exprese con rigor y que permita su manipulación de forma flexible. El lenguaje empleado tradicionalmente para representar el conocimiento es la lógica.

Una lógica se compone de una sintaxis y de una semántica. La sintaxis especifica qué axiomas están bien formados; la semántica, por su parte, define el significado de esos axiomas. En lógica clásica la semántica permite dos significados: verdadero o falso.

La lógica proposicional es la más simple, pero no por ello poco poderosa, de entre las utilizadas para inferir nuevo conocimiento. La sintaxis de la lógica proposicional incluye, entre otros, operadores como la negación (en inglés, *NOT*), la conjunción (*AND*), la disyunción (*OR*)... La semántica de la lógica proposicional, que determina si un axioma es verdadero o falso, se expresa mediante tablas conocidas, precisamente, como *tablas de verdad*. Para el caso de dos axiomas A y B, que pueden adoptar los valores binarios 0 (falso) y 1 (verdadero), la tabla de verdad correspondiente a los operadores básicos es:

A	B	NOT A	NOT B	A AND B
0	0	1	1	0
0	1	1	0	0
1	0	0	1	0
1	1	0	0	1

A OR B	A XOR B	A XNOR B	A NAND B	A NOR B
0	0	1	1	1
1	1	0	1	0
1	1	0	1	0
1	0	1	0	0

Es decir, que si A y B valen 0 (o, lo que es lo mismo, son falsos), entonces el resultado del operador *NOR*, que viene a significar «ni A ni B son verdaderos», dará 1 (verdadero). Pero dará 0, es decir, será falso, en cualquier otro caso.

La lógica proposicional permite representar hechos más complejos por medio de la combinación de diferentes operadores, por ejemplo: $A \wedge (B \vee C) \rightarrow D$ que debe entenderse como «si A es cierto y B o C son ciertos, entonces C también es cierto». Armados con esta lógica podremos deducir nuevo conocimiento a partir del conocimiento del que ya se dispone. La forma más básica de generar nuevo conocimiento en lógica proposicional es mediante reglas condicionales del estilo «si X entonces Y». Dos reglas de este tipo son las inferencias *modus ponens* y *modus tollens*. La inferencia *modus ponens* afirma que «si P implica Q, y P es cierto, entonces Q debe de ser cierto también». La inferencia *modus tollens* se define como «si P implica Q, y Q es falso, entonces P debe ser falso también».

Otras lógicas

Existen otras lógicas más sofisticadas que la proposicional para representar el mundo que queremos tratar. Se trata de lógicas capaces de recoger aspectos adicionales de la realidad aparte de simples he-

chos: por ejemplo, objetos y las relaciones que estos objetos guardan entre sí; o aspectos temporales de los hechos (en qué momento de una escala temporal tiene lugar) o diferentes grados de verdad sobre ellos (en qué porcentaje un determinado hecho es verdadero). Por ejemplo, la lógica de primer orden admite objetos (Sol, Tierra), hechos («La Tierra es un planeta», «El Sol es una estrella») y relaciones entre ellos («La Tierra gira alrededor del Sol», que se expresa como «gira[Tierra, Sol]»).

A la suma de los distintos aspectos de la realidad que un sistema lógico admite se le llama su *compromiso ontológico*, de ontología, la parte de la filosofía que estudia lo que hay. Otra característica importante de una lógica son los estados posibles de conocimiento sobre un hecho que puede manejar. En lógica proposicional, todo lo que se puede saber sobre un hecho es si es verdadero, falso o si se desconoce una cosa y otra. Otras lógicas, en cambio, admiten distintos grados de certidumbre acerca de un hecho. A la riqueza de conocimiento que una lógica es capaz de manejar se le denomina su *compromiso epistemológico*, de epistemología, la rama de la filosofía que se ocupa de lo que se conoce.

Una lógica que incluye la duda, como la *lógica difusa*, modela mucho mejor el razonamiento humano. Y es que el mundo real es tan complejo que no siempre todas las proposiciones son o verdaderas o falsas. Por ejemplo, en un determinado problema podemos necesitar expresar que llueve mucho o poco, y no simplemente si llueve. Imaginemos que debemos programar un asistente para que indique al usuario si debe llevar paraguas y/o usar abrigo cuando salga a la calle. Para ello debemos determinar bajo qué umbral la temperatura es lo suficientemente baja como para necesitar abrigo, así como determinar cuándo está lloviendo, diferenciando entre quizá una lluvia torrencial y unas gotas esporádicas. A través de la lógica difusa, las respuestas a reglas del tipo «si hace calor, no lleves abrigo» no serán «sí» o «no», sino una serie de grados en un rango entre 0 y 1. Frente a la pregunta de si debemos llevar paraguas ante

> GEORGE BOOLE Y LAS MATEMÁTICAS DE LA COMPUTACIÓN

— El pionero de la lógica computacional George Boole alrededor de 1860.

El álgebra de Boole es un sistema de reglas que permite tratar matemáticamente problemas lógicos de la forma «verdadero o falso», lo que a su vez hace posible expresarlos en forma de algoritmo y que un ordenador los entienda. Su nombre se debe a su inventor, el matemático británico George Boole (1815-1864). Las tres operaciones matemáticas básicas del álgebra de Boole son la negación o complemento, habitualmente escrita en inglés, *NOT*; la disyunción, *OR*; y la conjunción, *AND*. La negación, representada con el símbolo ¬, invierte el estado de una variable. Por ejemplo, si A es igual a «Sócrates es un hombre», entonces ¬A equivale a «Sócrates no es un hombre». La disyunción, representada con el símbolo v, es un operador binario, es decir, que necesita dos parámetros para obtener un resultado. Este es verdadero si alguno de los dos parámetros es verdadero. Por ejemplo, ¿es verdadero que lo que usted está haciendo ahora es viajar en un avión o leer? La respuesta es que sí, que es verdadero, y no importa que se encuentre cómodamente sentado en su sofá y no volando, porque es un hecho que usted está leyendo este libro. El tercer operador, la conjunción *AND*, equivaldría a reformular la pregunta anterior en la forma «¿es verdadero que lo que usted está haciendo ahora es leer y viajar en avión?» Si el operador le da 1, Boole y la buena educación nos aconsejan desearle un feliz vuelo.

una lluvia muy débil, quizá la respuesta sería «sí» en un 20%, y no «sí» taxativo La tabla siguiente recoge las ontologías y epistemologías de los sistemas lógicos más usuales:

Lenguaje	Compromiso ontológico	Compromiso epistemo-lógico
Lógica proposicional	Hechos	Verdadero / falso / desconocido
Lógica de primer orden	Hechos, objetos, relaciones	Verdadero / falso / desconocido
Lógica temporal	Hechos, objetos, relaciones, tiempos	Verdadero / falso / desconocido
Teoría de probabilidad	Hechos	Grado de certeza entre 0 y 1
Lógica difusa	Hechos con distintos grados de verdad comprendidos entre 0 y 1	Grado de verdad entre 0 y 1

Máquinas que deciden: los sistemas expertos

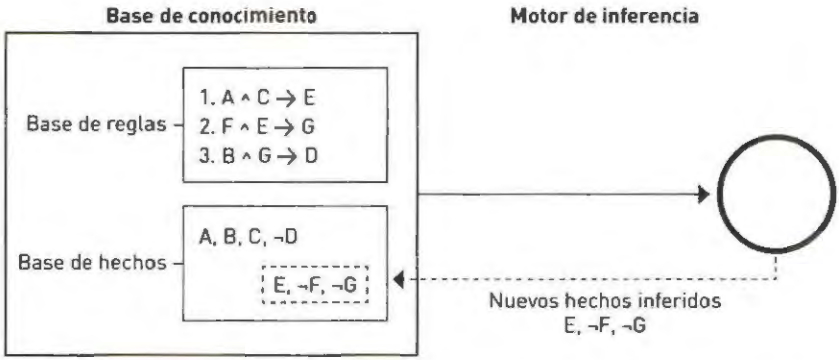
Un *sistema experto* emula el proceso de toma de decisión de un experto en un dominio específico. Muy en boga en los años 80 y 90 del siglo pasado, su número disminuyó cuando las plataformas estándar de software de negocios y gestión se apropiaron de buena parte de sus funcionalidades. Pero el término se ha continuado empleando para designar de forma genérica agentes basados en *software* capaces de llevar a cabo tareas diversas, aunque su tecnología tenga poco que ver con la que veremos en este apartado.

Las decisiones de un sistema experto se sustentan en una base de conocimiento formada por hechos y reglas. Esa base se enriquece con nuevo conocimiento generado por un motor de inferencia lógica que genera nuevos hechos. Por tanto, un sistema experto tiene dos componentes básicos: la *base de conocimiento*, por un lado, y el *motor de inferencia* o *sistema de razonamiento*, por el otro.

La versión más básica de un sistema experto funciona con lógica propositiva y razona basándose en los ya vistos *modus ponens* y *modus tollens*. Veamos un ejemplo donde los hechos A, B, y C son ciertos mientras que D es falso. Además, las reglas del sistema nos indican que: si A y C son ciertos, E también (regla 1); que si F y E son ciertos, G también lo es (regla 2); y que si B y G son ciertos, también lo es D (regla 3). A partir de estas reglas, el motor de inferencia deduce nuevos hechos: E es cierto, y F y G son falsos (fig. 3). La deducción funciona de la siguiente manera:

1. Aplicamos *modus ponens* a la regla 1 para deducir que E es cierto.
2. Aplicamos *modus tollens* a la regla 3 para deducir que o bien B o bien G son ciertos. Y como es un hecho que B es cierto, G ha de ser falso (lo cual se escribe formalmente como $\neg G$).
3. Aplicamos *modus tollens* a la regla 2 para deducir que o bien E o bien F son ciertos. Como hemos deducido anteriormente que E es cierto, F ha de ser necesariamente falso ($\neg F$).

Fig. 3



Esquema de cómo un sistema experto enriquece su conocimiento del mundo con nuevos hechos que se siguen de otros ya conocidos.

Una característica importante de los sistemas expertos es que son muy dinámicos: se puede añadir nuevo conocimiento sin necesidad de cambiar el código del programa. Esto es así porque su conocimiento (los hechos y las reglas) reside en un repositorio al margen del resto del sistema. Imaginemos un sistema experto para ayudar a los médicos a realizar diagnósticos y prescribir tratamientos. En la base de conocimiento tendríamos todas las reglas médicas que codifican el conocimiento médico. Las reglas serían del estilo de «si hay irritación de garganta, entonces el paciente tiene faringitis, laringitis o amigdalitis». Un sistema así sería muy amplio y flexible, porque para tenerlo al día bastaría con introducir nuevas reglas en la base de datos cada vez que la medicina adquiriera nuevos conocimientos. La introducción la podría llevar a cabo el médico directamente, lo que constituye otra ventaja. Dado un nuevo paciente, se introducirían todos sus síntomas y otras variables, y el motor de inferencia deduciría un nuevo hecho que, en este caso, sería el diagnóstico del paciente. El médico podría entonces revisar el diagnóstico y las reglas que se activaron para llegar a esa conclusión. Dichas reglas son claves porque ofrecen al médico la explicación de cómo se ha llegado a ese diagnóstico.

> La auténtica lógica del mundo es el cálculo de probabilidades.
>
> James Clerk Maxwell

A mayor flexibilidad, más expertos: las redes bayesianas

Los sistemas expertos basados en reglas de lógica proposicional como el que hemos examinado tratan problemas donde los hechos y las reglas son o verdaderos o falsos, sin otro matiz. Sin embargo, en la mayoría de ámbitos las decisiones se toman en situaciones de incertidumbre. Para reflejar esta circunstancia, hay sistemas

expertos cuyos hechos tienen asociada una probabilidad entre 0 y 1 de ser verdaderos. Estos sistemas se llaman *probabilísticos* o *estocásticos*. En el ejemplo del sistema experto médico anterior, por ejemplo, no se ha tenido en cuenta la incertidumbre, pero es un hecho que un mismo síntoma puede indicar diferentes enfermedades y con distinta probabilidad en cada caso.

Uno de los modelos más extendidos para simular el razonamiento en entornos de incertidumbre son las llamadas *redes bayesianas*. Las redes bayesianas se llaman así porque derivan nuevos hechos mediante *inferencia bayesiana*, un procedimiento que consiste en determinar cuán probable es un evento cualquiera A a partir de cuán probables son otros eventos relacionados con él. Por ejemplo, si se sabe que fumar es un factor causal de las enfermedades pulmonares, saber que un paciente fuma permite calcular, o al menos ajustar, la probabilidad de que padezca una enfermedad de esta clase. Para calcular este tipo de probabilidades condicionadas se emplea el llamado teorema o *regla de Bayes*. En términos algebraicos, esta regla se expresa como:

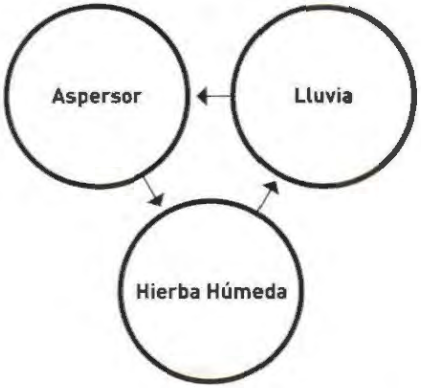
$$p(B|A)=\frac{p(A|B)p(B)}{p(A)}$$

Donde $p(B|A)$ es la probabilidad de que tenga lugar el evento B dado el A, $p(A|B)$ es la probabilidad de que tenga lugar el evento A dado el B, y $p(B)$ y $p(A)$ es la probabilidad de los eventos B y A, respectivamente.

Imaginemos, por ejemplo, que disponemos de dos variables que determinan que la hierba de un jardín esté húmeda: que el aspersor esté activado o que esté lloviendo (suponiendo que, si llueve, el aspersor se apaga en un 99% de las veces). Una vez disponemos de estas variables, podemos crear un modelo bayesiano en el que las tres variables tienen dos posibles valores (T verdadero, F falso) [fig. 4]. Las tres variables son: «hierba húmeda», «aspersor» y «lluvia».

Fig. 4

Lluvia	Aspersor T	Aspersor F
F	0,4	0,6
T	0,01	0,99



Lluvia T	Lluvia F
0,2	0,8

Aspersor	Lluvia	Hierba Húmeda T	Hierba Húmeda F
F	F	0	1
F	T	0,8	0,2
T	F	0,9	0,1
T	T	0,99	0,01

Un sencillo modelo bayesiano con tres variables relacionadas.

Mediante el teorema de Bayes, y asumiendo que la humedad de la hierba es el factor conocido, el modelo puede responder a preguntas del estilo de «¿cuál es la probabilidad de que esté lloviendo si la hierba está húmeda pero el aspersor está apagado?»

El modelo puede entenderse también como un grafo de tres nodos. De becho, una red bayesiana puede expresarse en forma de grafo donde los nodos representan las variables («aspersor», «hierba húmeda»), y las uniones entre nodos las influencias causales codificadas como probabilidades cruzadas («si el aspersor está encendido, la probabilidad de que esté lloviendo es del 0,01%»). Un nodo que sea padre de otro nodo, como «aspersor» lo es de «hierba húmeda», implica que es una causa directa del mismo.

Para responder a la pregunta anterior, basta con consultar la probabilidad de que «hierba húmeda» sea verdadero (T) si «lluvia» también lo es (T) pero «aspersor» es falso (F). El resultado es un 80%.

En el caso de diagnósticos médicos, una red bayesiana podría representar síntomas y enfermedades en sus nodos. Las uniones entre nodos expresarían las correlaciones entre síntomas y enfermedades obtenidas de estudios de casos médicos conocidos. A cada síntoma se le asigna, con respecto a una enfermedad con la que esté relacionada, un grado de sensibilidad y otro de especificidad. La *especificidad* del síntoma es la probabilidad de no tener el síntoma cuando la enfermedad está presente. La *sensibilidad* del síntoma, por su parte, es la probabilidad de que se aparezca cuando se padece la enfermedad relacionada. Los síntomas con un alto grado de sensibilidad y especificidad con respecto a una enfermedad son considerados como síntomas verdaderos de dicha enfermedad, y al contrario. Esta red sería capaz de responder preguntas del tipo «¿cuál es la probabilidad de que la aparición de un síntoma cualquiera denote una enfermedad concreta?»

Sistemas como este se están utilizando ya en hospitales para ayudar al médico a diagnosticar y proponer tratamientos médicos. Por ejemplo, el sistema de salud público de Inglaterra utiliza una red bayesiana para hacer un primer diagnóstico rápido. Naturalmente, las expuestas aquí son redes bayesianas de gran sencillez. Pero como estructuras básicas sobre las que edificar emuladores de la racionalidad humana ofrecen un gran potencial.

PERCEPCIÓN, COMUNICACIÓN E INTERACCIÓN

¿Qué aspectos son clave en nuestra interacción con el entorno? Obviamente los sentidos, en especial la vista. Pero también un cuerpo capaz de desplazarse y de manipular objetos. Y, desde luego, la posibilidad de comunicarnos con una parte esencial de ese entorno:

otras personas. De hecho, el dominio del lenguaje es, según algunos filósofos, la característica definitoria de la inteligencia humana. Turing, por ejemplo, hizo del dominio del lenguaje la condición suficiente de una máquina inteligente. Además, si tenemos en cuenta que buena parte de la información está en forma escrita o verbal, toda máquina que desee adquirir conocimiento de su entorno necesita comprender el lenguaje natural en sus distintas formas.

Modelos sencillos del lenguaje natural y sus funciones

Un lenguaje natural se define por medio de una gramática que especifica las reglas válidas de construcción de frases y de una semántica que define el significado de palabras, frases y textos. Sin embargo, la mayor parte del corpus escrito y, sobre todo, verbal incumple esas reglas en mayor o menor grado, por lo que los modelos computacionales más simples del lenguaje natural optan por tratarlo como inmensos conjuntos de cadenas de caracteres que aparecen con mayor o menor probabilidad. El más extendido de entre los modelos de este tipo es el denominado *n-gram*, que consiste en dar una probabilidad de ocurrencia a cada secuencia de *n* caracteres. Para ello, hay que disponer de un conjunto lo bastante extenso de documentos escritos en un idioma o corpus. En castellano, por ejemplo, la probabilidad de ocurrencia de la cadena de caracteres «hola» es de P(«hola») = 0,0034, mientras que P («adfg») = = 0,0000000000001. Los valores de cada una de estas probabilidades nos permiten concluir que «hola» es una secuencia de caracteres bien formada, y «adfg» no lo es. Los modelos n-gram se han demostrado muy útiles para identificar el idioma de un documento, dado que las probabilidades de ocurrencia de cadenas similares de caracteres son específicas de cada lengua. Incluso con solo con un par de palabras o tres en el documento este método permite identificar un idioma con un error menor del 0,5%.

Los modelos tipo n-gram también se pueden aplicar para la compresión de datos. Esta técnica detecta patrones en los textos y sustituyen dichos patrones por una representación más compacta. Por ejemplo, si la secuencia «por ejemplo» es detectada con cierta regularidad en un texto, se puede reemplazar por un número que es mucho más compacto que los 11 caracteres. Si esta técnica se aplica a todas las secuencias o patrones que aparecen en un texto con cierta asiduidad, tendremos un texto con la misma información pero mucho más compacto. Entre las funciones que los modelos n-gram desempeñan con mucha eficacia se encuentran, además del reconocimiento del idioma, estas: la corrección ortográfica; el filtrado de *spam*; el análisis de si una mención valorativa, por ejemplo, una crítica de cine, es positiva o negativa; y la detección de si un término o palabra clave está presente en un documento.

> Los límites de mi lenguaje son los límites de mi mundo
>
> LUDWIG WITTGENSTEIN

Esta última capacidad resulta fundamental a la hora de recuperar información de interés, como por ejemplo páginas *web* que traten de un tema concreto. Una opción en principio obvia para lograr este objetivo consistiría en diseñar un programa que, simplemente, seleccione las páginas donde las palabras de la consulta aparezcan más veces. Sin embargo, hay palabras populares cuya frecuencia de aparición es alta incluso en páginas donde no resultan especialmente relevantes. Para poder determinar la relevancia real de una palabra o palabras en una página es necesario tener en cuenta este efecto y descontarlo. Ahora bien, además de garantizar que las palabras solicitadas son relevantes en el contexto de una página, hay que asegurar que la página en sí también sea relevante. Pero ¿cómo es posible saber qué páginas son relevantes para quien hace la consulta solamente a partir de unas palabras clave? Dos estudiantes universitarios, Larry Page y Sergey Brin, decidieron emplear un criterio objetivo para determinar la relevancia de una página: sus

vínculos. En 1998, lanzaron al mercado Google un buscador que ordenaba las páginas seleccionadas de mayor a menor relevancia en función de ese criterio. Y el resto, como suele decirse, es historia.

El algoritmo que se encarga de ordenar las páginas en una búsqueda de Google se llama PageRank. El valor que PageRank otorga a una página es igual a la probabilidad teórica de que un navegador la visite aleatoriamente. Esa probabilidad es mayor cuantos más vínculos conduzcan a ella, y especialmente si esos vínculos proceden a su vez de páginas con mucho tráfico. La versión original y más básica del PageRank de una página u, $PR(u)$, obedece a la expresión:

$$PR(u) = \sum_{v \in Bu} \frac{PR(v)}{L(v)},$$

dónde Bu es el conjunto de todas las páginas que se enlazan con la página u, y $L(v)$ es el número de enlaces desde una página cualquiera v. De la fórmula se desprende que el PageRank de una página es la suma de los PageRank de las páginas que enlazan con ella, y que es recursiva.

Un paso más allá: entendiendo los lenguajes naturales

La utilidad de los modelos n-gram está limitada por sus necesidades de información. Así, para modelizar un léxico de 100 000 palabras (diccionarios representativos del inglés, el italiano y el japonés, por poner tres ejemplos, contienen 170 000, 270 000 y 500 000 palabras, respectivamente) se necesitarían 10^{15} cadenas de tres caracteres, es decir, que para obtener unas probabilidades de aparición representativas no bastaría siquiera con un corpus de mil billones de palabras. Cuando lo que se persigue es un objetivo tan ambicioso como modelar un idioma entero para servirse de él como herramienta de comunicación, hay que dar un paso más en la complejidad de los

modelos. Para ello, regresaremos a la idea inicial de gramática de un idioma como aquel conjunto de reglas que determina si una frase es válida en dicho idioma. También nos serviremos de nociones procedentes de la lingüística como las de categoría léxica, en la que se incluyen los sustantivos, los adjetivos, los verbos, etcétera, y los grupos sintácticos que resultan de agruparlos, como el sintagma nominal, el sintagma verbal y otros. A continuación, nuestro modelo establecerá reglas probabilísticas del tipo:

SV	→ Verbo (60%) → SV SN (40%)

Esta regla determina que un sintagma verbal (SV) está compuesto por un único verbo en un 60% de las veces y por un SV seguido de un sintagma nominal (SN) en el otro 40% (los porcentajes no son reales). A continuación, establecemos un diccionario de palabras admitidas, y a cada una le asignamos una probabilidad de aparición (la suma de probabilidades en cada categoría es 100%:

Sustantivo	Animal (0,05%), Barco (0,01%), Hámster (0,005%), Pedro (0,001%)...
Verbo	Es (10%), Está (1%), Parece (0,1%), Huele (0,001%)...
Adjetivo	Todo (1%), Muerto (0,1%), Verde (0,05%)...
Artículo	La (25%), Un (5%)...
Conjunciones	Y (20%), O (5%)...

Por último, listamos las reglas que constituyen nuestra gramática:

Regla 1

Frase	→ SN SV (90%) → Frase conjunción frase (10%)

Regla 2

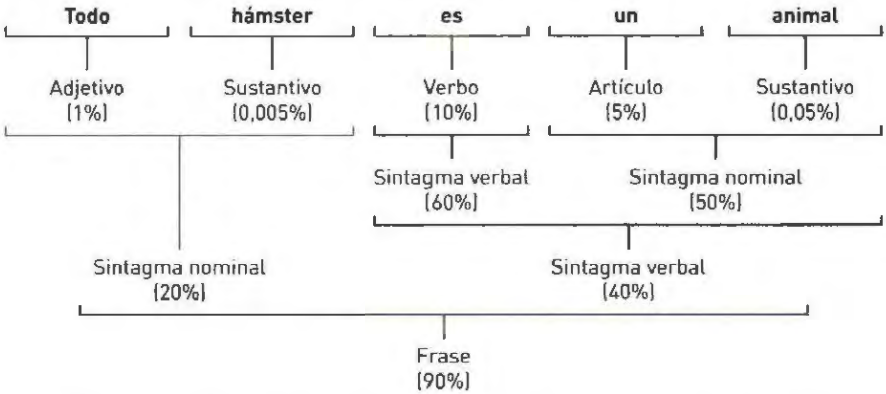
SV	→ Verbo (60%) → SV SN (40%)

Regla 3	
SN	→ Sustantivo (20%) → Artículo Sustantivo (50%) → Sustantivo Adjetivo (10%) → Adjetivo Sustantivo (1%)

Las reglas establecidas hasta ahora generan frases gramaticalmente correctas en castellano, como por ejemplo:

Todo hámster es un animal.
Pedro parece un hámster.

Las frases creadas por este sistema pueden analizarse mediante árboles sintácticos como este:



Como se observa, cada nodo tiene su probabilidad. La probabilidad total de esta frase concreta es 0,9 x 0,4 x 0,5 x 0,6 x 0,0005 x 0,05 x 0,01 x 0,2 x 0,00005 x 0,01, lo que arroja una cifra astronómicamente baja, como era de esperar tratándose de una sola frase en el contexto de todo un idioma. La ventaja de este tipo de árboles es que constituyen una prueba por construcción de que una frase cualquiera es válida según la gramática escogida. En el mundo real, la asignación de probabilidades de cada palabra y categoría sintáctica no se fija de antemano, como es obvio, sino que resulta del análisis de un corpus

real del idioma. Para realizar ese análisis, las frases del corpus tienen que estar todas en la forma de árbol sintáctico que acabamos de ver.

El paso siguiente es dotar a este modelo de una semántica, es decir, codificar no ya la estructura de una frase sino su significado. Los métodos de más éxito obedecen al principio de la semántica compositiva, según la cual el significado de una oración es una función del significado de las partes que la componen. En el caso de una frase sencilla, por ejemplo, «María quiere a Román», el SN «María» tiene como significado el término lógico *María*. Sin embargo, el SV «quiere a Román» no tiene una equivalencia lógica directa. Intuitivamente, vemos claro que se trata de una descripción que puede o no aplicarse a alguien o algo (en este caso, a María). Se trata, por tanto, de un predicado que, al combinarse con un nombre, da como resultado una frase lógicamente completa. Para reconstruir el significado de la frase estableceríamos una regla que dijera «un SN de significado *nombre* seguido de un VP de significado *predicado* da como resultado una frase cuyo significado lógico resulta de aplicar *predicado* a *nombre*», donde «aplicar *predicado* a *nombre*» es una operación previamente definida de la que resulta una notación lógica que el ordenador puede entender.

Esta exposición ha dejado de lado toda clase de particularidades del lenguaje que complican enormemente su modelización: los tiempos verbales, el tratamiento de palabras de significado contextual como «yo» u «hoy», la intención del hablante o la ambigüedad. Es por ello por lo que estas aproximaciones simbólicas a la comprensión del lenguaje natural se suelen complementar con otras de carácter estadístico.

La percepción artificial

La percepción artificial es la capacidad de una máquina de interpretar datos procedentes del entorno de una forma similar a cuan-

do los humanos interpretan las señales de sus sentidos. Básicamente, se distinguen tres tipos de percepción artificial: visión, oído y tacto. El enfoque general es el mismo para los diferentes tipos: recopilar datos a través de sensores para adquirir y relacionar información sobre el mundo que rodea a la máquina.

> No estamos escaneando todos esos libros para ser leídos por un humano, sino por un sistema de inteligencia artificial.
>
> Ingeniero anónimo de Google

Sin embargo, una representación perceptual no puede limitarse a ser una simple copia de la realidad, sino que se deben extraer ciertas propiedades útiles para el objetivo de la máquina. Además, lo que para nosotros pueden ser tareas naturales e incluso inconscientes, como la habilidad para centrarnos en una conversación con otra persona a pesar del ruido ambiente, para una máquina supone un desafío. De igual modo, percibir un objeto e identificarlo también acarrea complejos problemas teóricos y prácticos.

A la hora de percibir un estímulo cualquiera, el proceso debe descomponerse en diversas fases que se abordan por separado: captación (el dispositivo transductor transforma la información recibida a través de sensores en señales eléctricas), preproceso (la mejora de la información ya digitalizada a fin de, por ejemplo, eliminar el ruido de fondo), segmentación (separar el objeto que se quiere reconocer del resto de la información, como un rostro del resto de la fotografía), descripción (se representa el objeto con menos información y más relevante que el original, como la altura de los ojos y la posición de la nariz en un retrato), reconocimiento (saber qué objeto es o determinar algún detalle, como una grieta en una tuerca que se está examinando) y, finalmente, actuación (toma de decisión más o menos compleja).

En el caso de la visión, el área de investigación más importante en el campo de la percepción artificial, la fase más compleja es la del reconocimiento de imágenes. Hasta hace poco, los intentos de

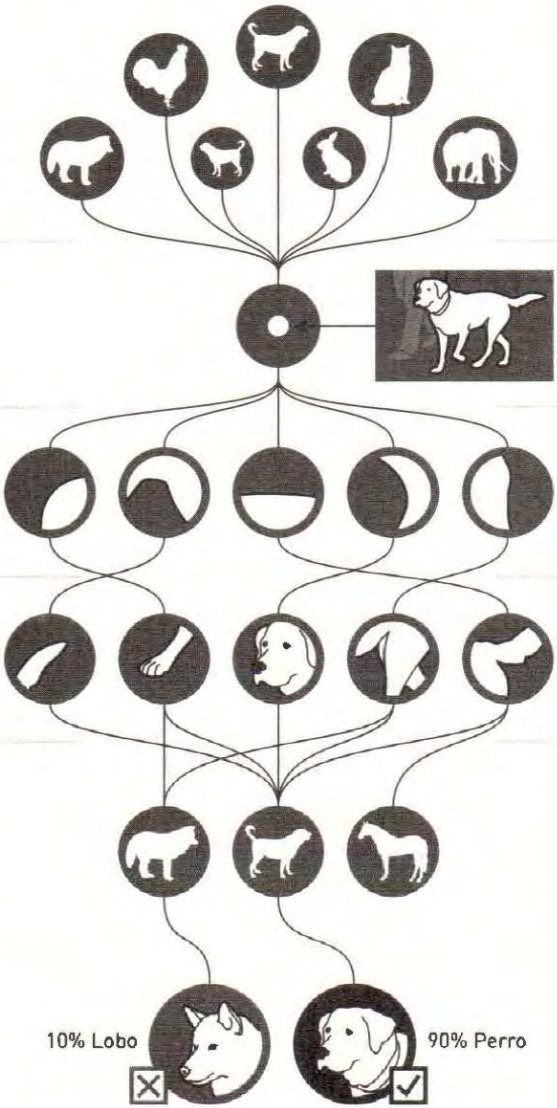
modelizar las habilidades humanas y animales en este ámbito se topaban con importantes obstáculos. De hecho, la visión artificial llegó a considerarse un problema «IA completo», es decir, que solucionarlo tenía una dificultad equivalente a la de conseguir una IA general. Ahora bien, como ha ocurrido con la traducción automática, las aproximaciones tradicionales de tipo simbólico se han visto rápidamente superadas por el desempeño de las redes neuronales. La enorme cantidad de datos visuales extraídos de internet y las redes sociales han permitido entrenar a redes neuronales que ya en 2015 superaron a los seres humanos en su precisión a la hora de reconocer una amplia gama de imágenes, desde razas concretas de perro a tipos de vegetación (fig. 5).

Robótica: cuerpos para las IA

Tanto la percepción artificial como la visión artificial son importantes complementos de la robótica, la disciplina de ingeniería mecánica, eléctrica y electrónica para el diseño, desarrollo y operación de robots. Un robot es una entidad mecánica artificial capaz de desarrollar una serie de tareas automáticamente. La IA entra en juego cuando esas tareas requieren inteligencia, como en el caso de un coche autoconducido.

Los logros de la IA tradicional en el campo de la robótica nunca alcanzaron el brillo de otros ámbitos. A finales de la década de 1980, el ingeniero australiano Rodney Brooks, miembro del laboratorio de IA del MIT, propuso un nuevo y revolucionario enfoque para la consecución de robots inteligentes. Brooks, como Herbert A. Simon antes que él, se percató de que el comportamiento de un insecto, como por ejemplo una hormiga, era tanto más complejo cuanto más le exigía el entorno con el que interaccionaba. Por tanto, propuso dejar de lado las habilidades cognitivas del robot y centrarse en que fueran capaces de desplazarse y reaccionar con el medio físico. De esta interac-

Fig. 5



Entrenamiento
Durante la fase de entrenamiento, una red neuronal es alimentada con miles de imágenes etiquetadas con diferentes animales, aprendiendo a clasificarlos.

Entrada
A esta red entrenada, le es mostrada una imagen sin etiqueta.

Primera capa
Las neuronas responden a diferentes formas básicas, tales como contornos.

Segunda capa
Las neuronas responden a estructuras más complicadas.

Última capa
Las neuronas responden a conceptos altamente complejos y abstractos que identificaríamos como animales diferentes.

Salida
La red predice qué objeto es el más probable, basado en su entrenamiento.

El presente esquema muestra cómo una red neuronal es capaz de reconocer un perro en una imagen en la que aparecen varios objetos sin identificar.

ción, confiaba Brooks, emergerían orgánicamente comportamientos inteligentes. Al igual que las hormigas, el robot llevaría a cabo una acción no por haberla planeado previamente, sino como respuesta al medio. Este enfoque pasó a llamarse *nouvelle AI*. De él resultó una escuela diferenciada en robótica conocida como *enfoque accionista* o también como *robótica basada en el comportamiento*.

En términos de los distintos modelos de agente que veíamos al principio del capítulo, la propuesta de Brooks equivale a un paso atrás: del agente basado en modelo al puramente reactivo. Pero hay que recordar que la investigación en IA atravesaba por aquella época su invierno particular, y que uno de los obstáculos principales que se le presentaban era cómo simular la percepción sin incorporar modelos computacionales a la postre inmanejables. Brooks apostó por saltarse un paso que todos creían fundamental, el de la cognición:

> Nada de cognición. Sólo detección y acción. Eso es todo lo que voy a construir, dejando completamente de lado lo que tradicionalmente se pensó como la inteligencia de la inteligencia artificial.

Brooks se propuso diseñar sus robots de forma que acciones complejas, como la exploración meticulosa de un entorno, se construían a partir de otra acción más elemental, como vagar sin rumbo por dicho entorno; esta, a su vez, de otra más sencilla todavía, y así sucesivamente. Brooks bautizó a este sistema de control como *arquitectura de subsunción*.

En 1988, uno de los primeros frutos de la *nouvelle AI*, Genghis, un robot caminador de seis patas que aprendía por sí mismo cómo moverse sobre tablas y otros obstáculos, causó sensación. Con el transcurrir de los años, esta clase de robots que aprenden del entorno se han ido sofisticando, como ejemplifica Baxter, un robot industrial desarrollado por Brooks en 2012. Dotado de brazos y una cara animada, Baxter puede usarse para tareas simples de carga, descarga, clasificación y manejo de componentes. Sin embargo, su

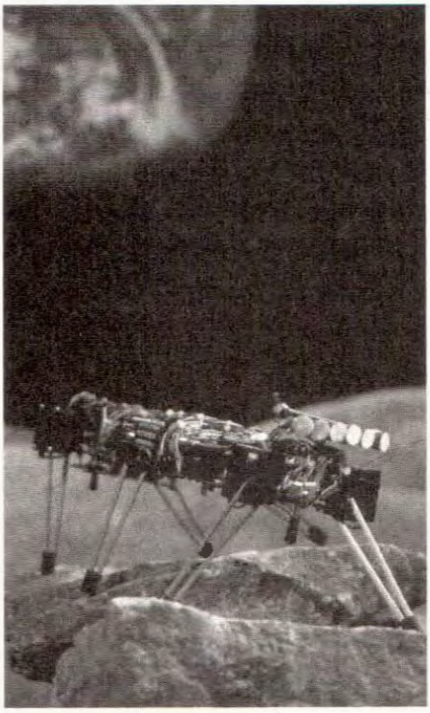
ventaja más destacable es que cualquier trabajador puede enseñarle las tareas que debe realizar, simplemente moviendo sus manos en la forma deseada. Baxter memoriza los movimientos y los repite cuando es necesario. Baxter, pues, no requiere de nueva programación, sino que aprende de forma intuitiva adaptándose al entorno.

Un camino a la IA general: la *embodied cognition*

La paradoja que se da entre la relativa facilidad de simular habilidades adultas complejas tales como la demostración de teoremas frente a la casi imposibilidad de hacer lo propio con tareas al alcance de un niño, como desplazarse por una habitación sorteando obstáculos, ya fue advertida por Hans Moravec. La razón de esta aparente paradoja reside en que las habilidades motoras y perceptivas del ser humano son producto de millones de años de selección evolutiva, por lo que se desempeñan a un nivel inconsciente. Sin embargo, si aspiramos a emular cualquiera de estas habilidades a través de ingeniería inversa, exigirán un nivel de dificultad proporcional al tiempo que han tardado en evolucionar. En otras palabras: las habilidades que nos resultan más sencillas serán las más difíciles de emular, y las que nos resultan más complejas (matemáticas, lógica, ingeniería), las que menos, porque han tenido unos pocos miles de años para ser refinados, principalmente por la evolución cultural.

Si la percepción y el movimiento, dos habilidades que implican un cuerpo físico y unos sentidos, están tan imbricadas en nosotros como para resultar inconscientes, podría ser que también desempeñaran un papel necesario en otros procesos mentales como la cognición. Esa es precisamente la tesis defendida por el movimiento de la *embodied cognition* o cognición corporal.

La cognición corporal defiende que la tradicional dualidad entre cerebro y cuerpo, que reserva al primero las tareas cognitivas y relega al segundo a mero canal de entrada y salida de datos sensoriales, es



— Tres creaciones de Rodney Brooks. Arriba, el ingeniero australiano con Baxter. Abajo, a la izquierda, Genghis, un sonado éxito de la robótica basada en el comportamiento. Sobre estas líneas, Cog, un robot creado en 1993 que aprendía mediante la interacción con humanos.

radicalmente falsa, y que tanto el cuerpo como el cerebro están implicados en la cognición. Así, se ha comprobado que, cuando se planea agarrar un objeto, en lugar de, por ejemplo, señalarlo, el cerebro ubica dicho objeto en el espacio de forma distinta. Es decir, que la acción corporal prevista altera el proceso cognitivo previo. También se ha comprobado que al representar físicamente una historia, el cerebro la retiene mucho mejor. Algunos investigadores incluso han afirmado que para comprender conceptos abstractos tales como la muerte, los lazos familiares o ciertos términos matemáticos necesitamos haber interiorizado previamente acciones físicas como medir distancias o estimar tamaños visualmente.

Si se asumen las premisas de la cognición corporal, los robots serían una vía necesaria para alcanzarla. Se trata de una idea con un buen número de partidarios entre los expertos, como pone en evidencia que el 35% de ellos señalara las *mentes corporeizadas* como un factor que iba a ser decisivo en el desarrollo de una IA general en la encuesta del primer capítulo.

Aunque el modelo de la cognición corporal fuera erróneo, la construcción de robots autónomos sigue siendo un ámbito relevante en IA. Y no solo porque gracias a nuestras pruebas con ellos podemos obtener importantes datos sobre los organismos vivos, sino porque está demostrado que los seres humanos se sienten mucho más cómodos atribuyendo inteligencia a entes antropomórficos.

04 MÁQUINAS QUE APRENDEN

La acumulación de nuevo conocimiento es una de las habilidades cognitivas fundamentales del ser humano.
En la ambiciosa tarea de simularla, una tecnología destaca sobre el resto: las redes neuronales.

¿Cómo aprende un niño que los vasos de cristal pueden romperse, pero los de plástico no? Naturalmente, de la experiencia. Los niños ven cómo ciertos objetos se rompen al caer al suelo, a diferencia de otros similares que no lo hacen. Tras observar varias veces el mismo fenómeno, comprenden que los primeros están hechos de un material distinto que los segundos (un material que, además, es transparente). Sería desde luego posible programar un ordenador para que identificara qué vasos se rompen, por ser de cristal, y cuáles no. Para ello habría que indicarle una serie de principios generales, como, por ejemplo, qué distingue a los vasos de otros objetos, y al cristal y al plástico de otros materiales. Si lo dotamos de un motor de inferencia, el ordenador podrá entonces concluir por sí solo que los vasos de cristal, pero no los de plástico, se rompen. Este sería el enfoque tradicional, a menudo llamado *determinista* dado que requiere que especifiquemos o determinemos de antemano todas las opciones relevantes. El niño opera de forma distinta. En lugar de llegar a una conclusión concreta («los vasos de cristal se rompen») a partir de otras más generales («los objetos de cristal

se rompen», «los vasos son objetos», etc.), hace la operación inversa, es decir, llega a un principio general («los vasos de cristal se rompen») a partir de observaciones concretas de objetos que caen al suelo y se rompen. Este segundo modo de aprendizaje, el que hemos atribuido a los niños, se llama *inducción*. El de los ordenadores deterministas, por su parte, se conoce como *deducción*. La inducción tiene la enorme ventaja de que no necesita principios generales, o solo unos pocos. Ahora bien, sí necesita de ejemplos concretos, en ocasiones gran cantidad de ellos, para llegar a conclusiones. Además, dichas conclusiones no pueden considerarse ciertas absolutamente. A diferencia de la deducción, que obtiene certezas absolutas, la inducción solo puede afirmar que algo es cierto con mucha probabilidad. Ahora bien, ¿cuál de los dos tipos de creencia nos parece más propio de los seres humanos? ¿Las creencias absolutas e inmutables, o las que cambian en función de lo que el mundo va mostrando? ¿Y cuál está más preparado para adaptarse a un entorno cambiante? Sin duda, la inducción se acerca más a nuestra experiencia y es más adaptable.

> El cerebro del niño es, seguramente, como un bloc nuevo. Un mecanismo sencillo y un montón de hojas en blanco.
>
> ALAN TURING

El *aprendizaje automático*, un término adaptado del inglés *machine learning* («aprendizaje de máquinas»), aspira a crear máquinas que aprendan como lo hacen los niños, es decir, por la inducción de principios generales a partir de observaciones concretas de hechos de su entorno. Una IA capaz de aprendizaje automático funcionaría de forma totalmente diferente a la de un ordenador determinista. En el caso de los vasos, por ejemplo, crearíamos un conjunto de experiencias de miles, decenas de miles o incluso millones de vasos que se caen al suelo. La IA analizaría todas estas experiencias y concluiría que los vasos hechos de cristal se rompen.

Naturalmente, los niños no aprenden solo mediante la observación. También reciben información de tipo directo. Por ejem-

plo, en una visita al zoo, el niño contempla un leopardo, y exclama «¡Mira, un gato!». Su padre o su madre le dicen que no, que no es un gato, sino un leopardo. En este caso, el niño ha aprendido más cosas sobre gatos (porque sabe que hay cosas similares que no son gatos) y empieza a aprender lo que es un leopardo. O sea, que el niño aprende de objetos o experiencias que le vienen, en cierto modo, etiquetados.

El aprendizaje a través de ejemplos etiquetados previamente se llama *supervisado*. Formalmente, el aprendizaje supervisado se define como:

> Dado un conjunto de entrenamiento de N ejemplos definidos como parejas de datos de entrada-salida (x_1, y_1), (x_2, y_2), ...(x_N, y_N), donde y_i fue generado con una función desconocida $y = f(x)$, descubrir una función h que aproxime la función desconocida f.

Un dato de entrada x_j podría ser, por ejemplo, una imagen, y el dato de salida y_j esa misma imagen con la etiqueta «es un gato» o la contraria. La función h es la hipótesis. Un algoritmo de aprendizaje supervisado busca la hipótesis que mejor explica los datos de salida. Una vez se cree que se ha alcanzado una buena explicación, se aplica a ejemplos no etiquetados. Si la función h es capaz de etiquetarlos correctamente, concluiremos que funcionará bien para cualquier ejemplo nuevo: el algoritmo ha aprendido a reconocer gatos.

Según el tipo de información, los algoritmos serán unos u otros. Cuando el dato de salida es un valor extraído de un conjunto, como, por ejemplo el estado del cielo en un día cualquiera (soleado, nuboso, lluvioso, etc.), se dice que el problema es de clasificación; si el conjunto está formado únicamente por dos valores (sí y no, verdadero y falso), se dice que es de *clasificación binaria*. En el caso de que las salidas sean numéricas (la temperatura de un día cualquiera, o el precio de un objeto), el problema de aprendizaje es un problema de *regresión*.

ÁRBOLES DE DECISIÓN Y MODELOS DE REGRESIÓN

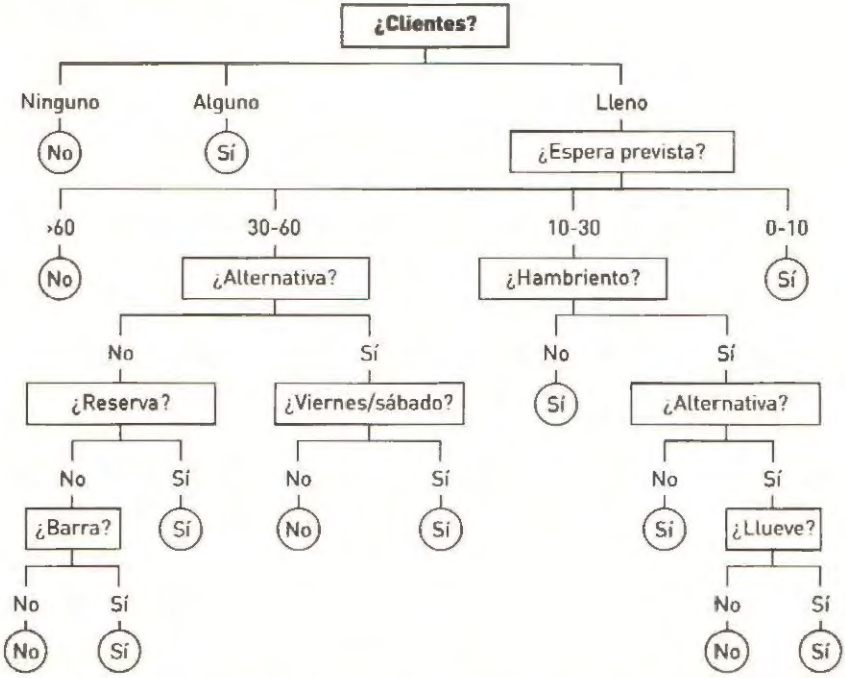
Una de las técnicas más sencillas e intuitivas para resolver mediante aprendizaje automático supervisado un problema de clasificación es la de los *árboles de decisión*. Para definir el problema a tratar hay que decidir cuáles van a ser las variables o *atributos* relevantes a la hora de tomar la decisión. Imaginemos un problema de clasificación binaria que consiste en saber si vale la pena esperar para tener mesa en un restaurante. Los atributos relevantes a la hora de tomar la decisión podrían ser los ocho siguientes:

¿Hay alternativa?	¿Tiene barra?	¿Es viernes o sábado?	¿Estoy hambriento?
¿Hay clientes?	¿Llueve?	¿Me reservan plaza?	¿Espera prevista?

Los datos de entrenamiento, expresados en función de los atributos seleccionados, podrían ser:

Ejemplo	*Restaurante*$_1$	*Restaurante*$_2$	*Restaurante*$_3$	*Restaurante*$_4$
¿Hay alternativa?	Sí	Sí	No	No
¿Tiene barra?	No	No	Sí	Sí
¿Es viernes o sábado?	No	No	No	No
¿Estoy hambriento?	Sí	Sí	No	No
¿Hay clientes?	Ninguno	Lleno	Algunos	Algunos
¿Llueve?	No	No	No	No
¿Me reservan plaza?	Sí	No	No	No
¿Espera prevista?	>60	30-60	0-10	10-30
Valor salida (¿Esperar?)	No	No	Sí	Sí

Una vez introducidos los datos en nuestro algoritmo, se generaría el árbol de decisión siguiente:

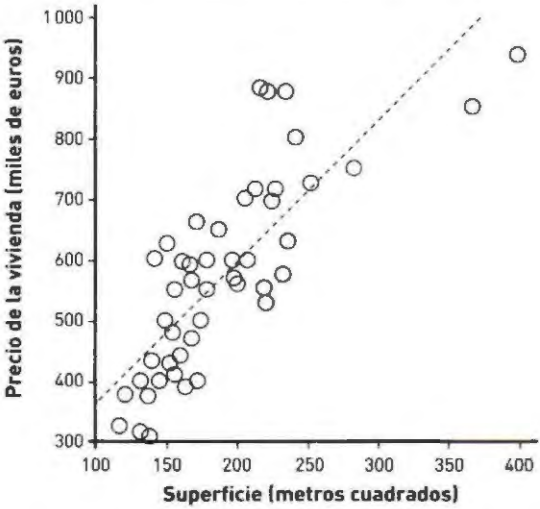


Como puede verse, los nodos del árbol representan atributos, las ramas los valores posibles de dichos atributos y las hojas los valores de salida. Cada conjunto de ramas desde la raíz hasta las hojas representa una regla posible de clasificación.

Uno de los peligros en los que se puede caer a la hora de generar un árbol de decisión es el de sobrealimentarlo con un exceso de datos de entrenamiento. En este caso, el árbol funcionará muy bien para dicho conjunto pero tendrá problemas para generalizar lo aprendido.

La regresión, por su parte, es un método estadístico para encontrar la relación que pueda existir entre una variable que nos interesa explicar y otras que creemos tienen influencia sobre ella. Supongamos que queremos averiguar si la superficie de un piso explica su

precio, y hasta qué punto. Para ello, marcamos con un círculo la superficie y el precio de varios pisos y los disponemos en una gráfica:



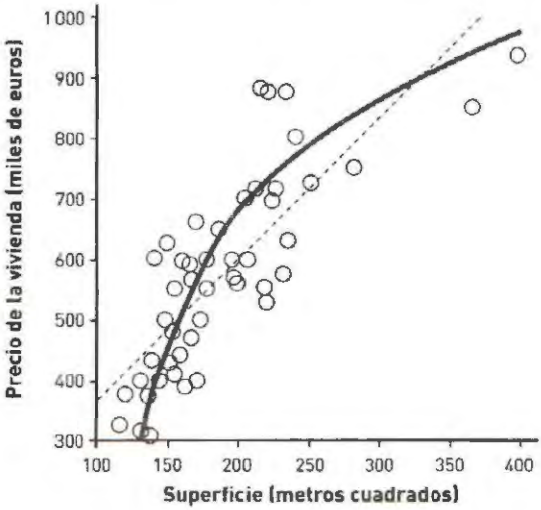
Ya a simple vista parece obvio que hay una relación positiva y creciente entre superficie y precio, y que dicha relación seguiría, más o menos, una recta como la dibujada. El método de la *regresión lineal* nos permite comprobar si nuestra intuición es correcta y darle a la relación una forma matemáticamente exacta.

En nuestro ejemplo de los pisos, al incluir solo una posible variable explicativa x, la superficie, nos hallamos ante una regresión lineal *simple*. Sin embargo, se pueden incorporar en el análisis tantas variables explicativas adicionales como creamos que ayudan a predecir con mayor detalle el dato que nos interesa. Está claro que la inclusión de factores adicionales como la ubicación o la antigüedad del piso mejoraría el valor predictivo de nuestro modelo.

La regresión lineal también puede usarse para resolver problemas de clasificación binaria. En este caso, el objetivo es encontrar la función que separe los elementos en dos clases.

La última gran familia de algoritmos de regresión es la *regresión no lineal*. En una regresión no lineal, la relación entre la variable que

se quiere explicar y las que supuestamente la explican no es una recta, sino una curva. Si aplicáramos un algoritmo de regresión no lineal a nuestro ejemplo, obtendríamos una curva parecida a la siguiente:



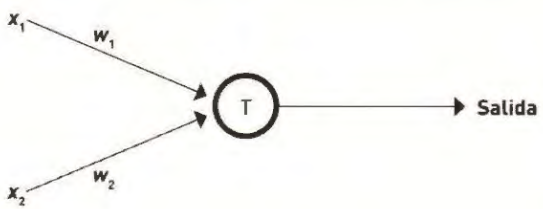
El resultado es un modelo que, tal vez, se adapta mejor a la realidad. El precio que hay que pagar tiene una mayor complejidad técnica, por lo que en ocasiones es una mejora de la que se puede prescindir.

LA GRAN PROMESA DE LAS REDES NEURONALES

A pesar de la gran utilidad de los árboles de decisión y de las regresiones, ninguna de estas técnicas tiene la versatilidad de las redes neuronales. Entre sus virtudes básicas se hallan la capacidad de lidiar con datos inadecuados o corruptos, y la posibilidad de distribuir la fase de entrenamiento en varias unidades de proceso para ganar rapidez y capacidad. Las redes neuronales son programas que intentan emular el funcionamiento del cerebro humano a escala neuronal. Vamos a repasar sus principios fundamentales. Las redes neuronales están compuestas por unidades neuronales conec-

tadas entre sí. Una neurona artificial posee una o más conexiones de entrada y una de salida. En cuanto recibe un valor de entrada, la neurona lo multiplica por un peso y, si el resultado supera un determinado umbral, propaga el valor inicial a las neuronas de las capas siguientes. En caso contrario, lo inhibe. A las reglas que determinan si una señal se propaga o no se las llama *función de conexión*.

A pesar de la sencillez de su planteamiento, nuestro ya conocido perceptrón, es capaz de llevar a cabo las operaciones lógicas fundamentales AND, OR y NOT. Tomemos como ejemplo la disyunción, OR:



x_1	x_2	x_1 OR x_2
0	0	0
0	1	1
1	0	1
1	1	1

¿Hay unos valores de los pesos w_1, w_2 y del umbral T de forma que para cada valor de entrada x_1, x_2 el valor de salida sea el predicho por la tabla? El valor de salida de un perceptrón se calcula:

$$\text{Salida} = \begin{cases} 0 & \text{si } \sum_j w_j x_j \leq \text{Umbral} \\ 1 & \text{si } \sum_j w_j x_j > \text{Umbral} \end{cases}$$

Si damos a todos los pesos y al umbral el valor 1, el perceptrón devuelve el valor deseado:

x_1	x_2	w_1	w_2	$x_1 w_1$	$x_2 w_2$	$x_1 w_1 + x_2 w_2$	T	Salida
0	0	1	1	0	0	0	1	0
0	1	1	1	0	1	1	1	1
1	0	1	1	1	0	1	1	1
1	1	1	1	1	1	2	1	1

Ahora bien, cuando intentamos computar operaciones lógicas más complejas, como la XOR, nuestro perceptrón es insuficiente.

Este fue, precisamente, el problema detectado por Minsky y Papert que veíamos en el segundo capítulo, y la solución, como ya dijimos, consistió en añadir capas adicionales de neuronas.

Aunque la red neuronal multicapa resultante puede, ahora sí, computar toda clase de operadores, tiene importantes limitaciones a la hora de aprender. El problema radica en que los perceptrones solo trabajan con valores binarios, 0 o 1, tanto de entrada como de salida. Recordemos que una red neuronal aprende en la medida en que es capaz de ajustar progresivamente los pesos y el umbral de activación hasta dar con la combinación que resuelve el problema planteado de forma fiable. Este ajuste se produce «hacia atrás», en el sentido de que la salida se compara con la salida esperada y la diferencia o error se propagan por la red neural en sentido contrario al habitual: de la salida a la entrada. A medida que se introducen cada vez más datos de entrenamiento, la red neuronal va ajustando de forma automática sus parámetros hasta ser capaz ella misma de proporcionar a los ejemplos resueltos la solución correcta. Una vez alcanzado ese punto, la red neuronal es ya capaz de dar respuesta a ejemplos no resueltos.

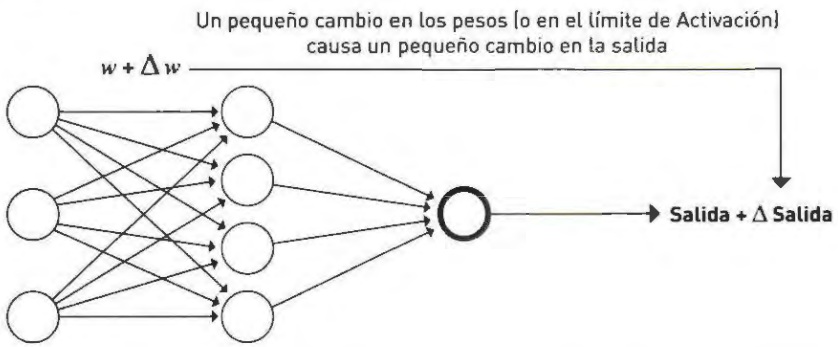
Para poder llevar a cabo esta tarea de aprendizaje es necesario que se cumpla la propiedad de que una alteración pequeña en los parámetros produzca también una alteración pequeña en la salida (fig. 1). De no ser así, los cambios necesarios para que una salida sea correcta harían que en otro punto de la red el resultado no fuera el deseado.

En una red neuronal de perceptrones, los cambios de pesos y lindares de activación pueden hacer cambiar por completo la salida de 0 a 1 (o de 1 a 0), y el ajuste gradual necesario, por tanto, es imposible.

Esta limitación se puede superar con otro tipo de neuronas con una función de activación más compleja que la superación de un umbral de activación. En tal caso, las salidas no son solo 0 o 1 sino que pueden tomar cualquier valor entre 0 y 1. Así ya es posible modular la magnitud de los cambios en los parámetros y los valores de salida.

Sobre este diseño básico caben toda clase de sofisticaciones, como en el caso de las redes neuronales *convolucionales*, empleadas para

Fig. 1



El aprendizaje mediante retropropagación requiere que los ajustes de los pesos o del umbral, y por consiguiente en la salida, sean modulables.

el reconocimiento y la clasificación de imágenes. Este tipo de redes neuronales se inspiran en el funcionamiento del córtex visual del cerebro biológico. Cada una de sus capas tiene un propósito específico: las hay convolucionales, de reducción de muestreo y de clasificación. Una imagen que entra en una capa convolucional emerge de ella filtrada y alterada de forma que ciertas características dominantes ya presentes en la imagen de entrada salen reforzadas. Las capas de reducción de muestreo, a su vez, simplifican la imagen reduciendo su resolución. De la combinación de unas y otras se obtiene una versión de la imagen expresada en términos de las características relevantes para la clasificación. En la capa de salida, finalmente, las imágenes son clasificadas según esos características (fig. 2).

La última frontera del aprendizaje: el *deep learning*

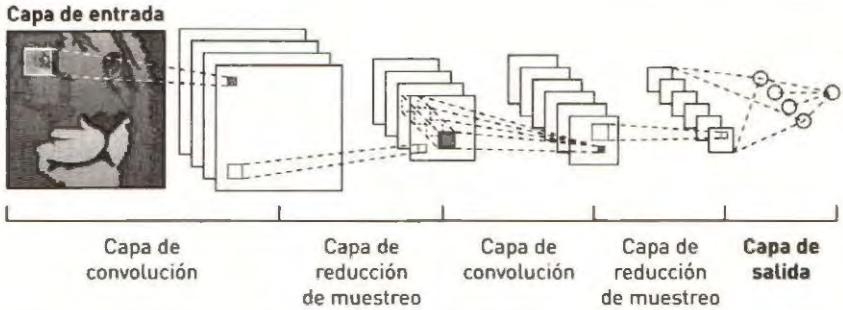
El *deep learning*, término inglés para «aprendizaje profundo», es un tipo de algoritmo de aprendizaje diseñado para aprovechar al

máximo la capacidad de procesar información en paralelo de las redes neuronales (o, también, de las redes bayesianas que hemos visto anteriormente). La idea subyacente es que el mejor modo de analizar los datos procedentes del mundo real es descomponerlos en distintos niveles de abstracción. En un contexto de redes neuronales, cada uno de estos niveles se correspondería con una red neuronal individual de tal forma que la salida de la red de menor abstracción sería la entrada de la de abstracción inmediatamente superior. Por tanto, cuantas más redes, mayores niveles de abstracción a la hora de descomponer los datos.

En el caso de un algoritmo de *deep learning* aplicado al reconocimiento de imágenes en un vídeo, la primera capa podría procesar el sonido, la segunda podría analizar la forma del objeto, la siguiente podría clasificar el objeto en diferentes formas básicas, etc. Cada paso a la siguiente capa de neuronas genera un razonamiento cada vez más abstracto.

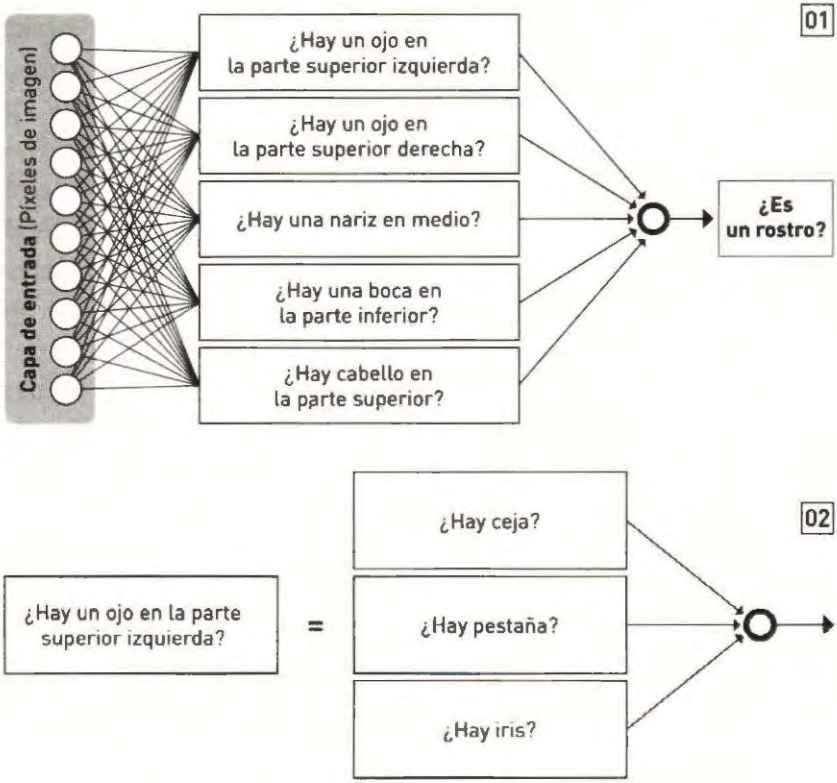
Para reconocer con técnicas de *deep learning* si en una imagen hay un rostro humano, por ejemplo, desglosaríamos el problema

Fig. 2



El campo visual es fruto de la suma de los inputs parciales de las neuronas del córtex visual. Una red convolucional se inspira en este principio para, mediante varias iteraciones del proceso, reducir una imagen a alguno de sus subcomponentes. Las imágenes son finalmente clasificadas en función de ese subcomponente.

Fig. 3



En este ejemplo de *deep learning* aplicado a una red neuronal, la tarea de clasificar imágenes como correspondientes o no a un rostro humano (**01**) se descompone en sucesivas capas de abstracción creciente (**02**), cada una de las cuales ocupa una red neuronal propia.

en diferentes abstracciones (o capas). En primer lugar, definiríamos un rostro como una imagen compuesta por dos ojos, una nariz, una boca y cabello ubicados de una forma determinada. A su vez, la identificación de un ojo la desglosaríamos en subtareas como la detección de una ceja, una pestaña, un iris, etc. (fig. 3). Esta jerarquía de mayor a menor abstracción podría hacerse tan profunda como se requiera.

La tecnología del *deep learning* ha experimentado un auge extraordinario en tiempos recientes a raíz de la evolución de la capacidad de cómputo de los ordenadores. Los algoritmos de *deep learning* están detrás de los grandes avances realizados en reconocimiento de voz y procesamiento del lenguaje natural que cimientan muchas de las aplicaciones estrella de la IA.

Con todas sus ventajas, las redes neuronales no son perfectas. Es muy difícil, por ejemplo, aprovecharlas para realizar otra labor que la inicialmente programada. Sin embargo, el inconveniente más importante es el de su opacidad. Un algoritmo de búsqueda tradicional como los que hemos visto es transparente acerca de cómo alcanza sus respuestas. Las redes neuronales, en cambio, son como una caja negra de la que es imposible extraer información sobre qué las ha llevado a una solución. La opacidad de las redes neuronales nos sitúa de nuevo ante el dilema que planteó Searle sobre la inteligencia. ¿Es la inteligencia algo más que comportamiento inteligente?

En el ámbito de IA débil, donde las cuestiones filosóficas tienen poca o nula importancia, las redes neuronales y las técnicas que se sirven de ellas, como el *deep learning*, han experimentado un gran auge. Desde programas que pasan, mejor que cualquier humano, pruebas de acceso a la universidad tan exigentes como las japonesas, hasta otros capaces de generar obras pictóricas originales a partir del estudio sistemático de las pautas de grandes maestros como Rembrandt, las redes neuronales nos están descubriendo pautas insospechadas en los datos. Ya ha llegado el día en el que sistemas de este tipo analizan nuestro comportamiento en cuanto consumidores y generan a partir de las regularidades detectadas recomendaciones tan poco obvias como acertadas. En teoría, esta técnica podría ampliarse a otros ámbitos de nuestro comportamiento, como el amoroso. Tal vez llegue el día en el que dejaremos que un programa tome por nosotros decisiones personales vitales confiados en su conocimiento profundo de nuestros gustos e inquietudes, en especial de aquellos de los que no somos conscientes.

APRENDIZAJE SIN SUPERVISIÓN

¿Qué ocurre cuando se quieren establecer patrones de comportamiento pero no se dispone de un histórico de datos que sirvan como ejemplo? En estos casos, se acude a técnicas de aprendizaje en IA capaces de agrupar datos según características comunes sin necesidad de entrenamiento ni supervisión. La noción clave en el *aprendizaje no supervisado* es la de distancia entre datos. En un algoritmo de los llamados de *agrupación jerárquica*, por ejemplo, un grupo o *cluster* se define por la distancia mínima necesaria para incluir los objetos que le pertenecen. Los algoritmos *k-means*, por su parte, hacen justamente la operación contraria. En lugar de tener fijada una distancia y ver qué agrupaciones resultan de ella, encuentra las distancias que definen un número *k* de grupos decididos de antemano.

El aprendizaje no supervisado se utiliza en *marketing* digital, por ejemplo, para descubrir patrones o agrupaciones en grupos de consumidores sobre los que no se dispone de datos. Imaginemos el caso de un banco que tiene la información de todas las compras realizadas con cargo a sus cuentas en una ciudad. Un sistema de aprendizaje no supervisado podría, basándose en esa información, agrupar a los consumidores por niveles de gasto y otras regularidades en las compras. En el *marketing* tradicional, es el profesional quien decide qué características del comportamiento de compra del consumidor son útiles para poderlos diferenciar. En un paradigma basado en el dato como el que ejemplifica el aprendizaje automático, es la máquina quien «decide» o en el peor de los casos «propone» al experto humano una serie de agrupaciones posibles. La ventaja obvia de este sistema es que no es víctima de las ideas preconcebidas típicas de la empresa o el sector y, por tanto, puede hacer aflorar relaciones insospechadas y de gran valor. Otra ventaja es que puede tomar decisiones automáticas en función de las regularidades detectadas en tiempo real, lo que resulta clave a la

hora de cerrar una venta impulsiva o en entornos digitales de venta como el *e-commerce*.

Otro tipo de problemas en los que el aprendizaje no supervisado ofrece un valor considerable es el de la detección de anomalías. En un caso así, el sistema detecta las características habituales de un grupo de datos y, cuando un dato se aleja de ese patrón de normalidad, queda señalado como anomalía. Esta información puede ser clave a la hora de, por ejemplo, detectar al instante una transacción financiera fraudulenta entre millones de ellas.

JUNTANDO TODAS LAS PIEZAS: LAS ARQUITECTURAS COGNITIVAS

Además de su valor intrínseco, el aprendizaje es un elemento fundamental de todo proceso cognitivo. Por tanto, es de esperar que cualquier IA general incorpore esa habilidad. Y lo mismo cabe decir del razonamiento y de la interacción con el entorno. En el arranque del tercer capítulo proponíamos la noción de *agente racional* y afirmábamos que su diseño era el fin último de la IA. Tras el recorrido que hemos hecho por los ámbitos de investigación y las tecnologías más destacadas de la disciplina, podemos concretar algo más en qué podría consistir un agente racional. La respuesta nos retrotrae a la encuesta citada en el primer capítulo, y al enfoque que, según los expertos, era el más prometedor a la hora de conducirnos al Santo Grial de la IA humana: las *arquitecturas cognitivas*.

Las arquitecturas cognitivas proponen un modelo computacional de la cognición, inspirado en la mente humana, que encapsule en un solo proceso la racionalidad, la acción y el aprendizaje, principalmente; pero también la reflexión, la memoria, etcétera. El término «arquitectura» implica que se intenta modelar no solo el comportamiento, sino también las propiedades estructurales del sistema analizado. Las arquitecturas cognitivas suelen basarse en

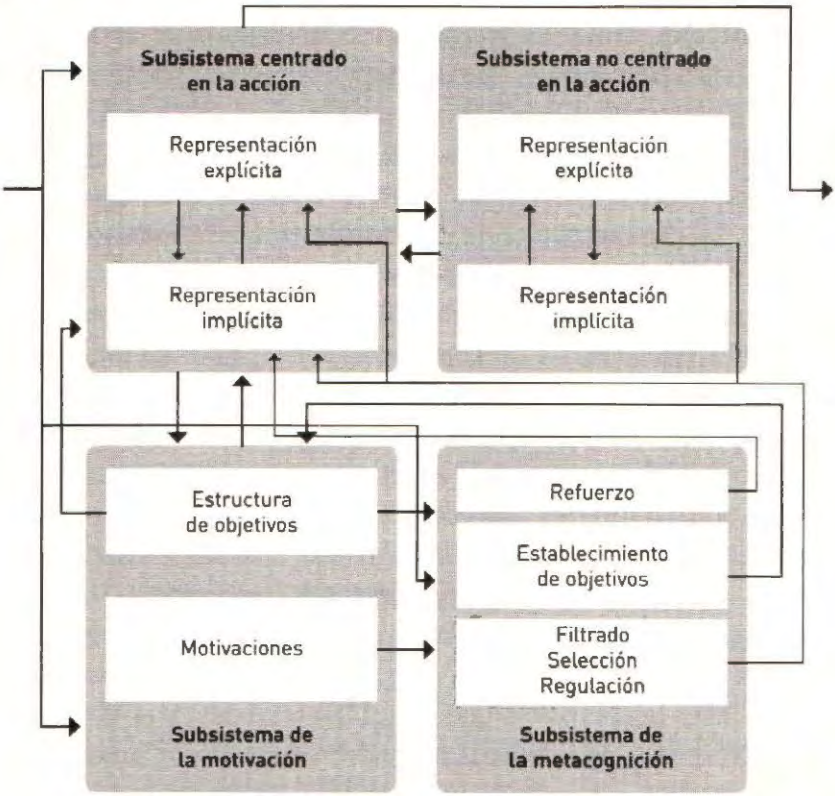
hipótesis creíbles sobre cómo opera nuestra mente, según las evidencias procedentes de la psicología y la ciencia cognitiva. Existen muchos modelos diferentes de arquitecturas cognitivas, principalmente en función de qué enfoque privilegian a la hora de modelizar la cognición: el simbólico, el conexionista o un híbrido entre ambos.

Tal vez la más destacada de las encuadradas en el enfoque simbólico, la arquitectura Soar, creada en 1983 por iniciativa de John Laird, Paul Rosenbloom y nuestro viejo conocido Allen Newell, tiene como objetivo desarrollar los «bloques computacionales fijos constitutivos de un agente inteligente». En el modelo Soar, esos bloques o piezas se combinan entre sí con el objetivo último de emular al ser humano en todas las tareas que impliquen cognición. Es, por tanto, un modelo cuya ambición no es otra que la IA general. Uno de los principios que cimientan la arquitectura Soar es que toda tarea asociada a un objetivo puede representarse como un problema de búsqueda, del estilo de los que vimos en el capítulo tercero. También comparte con otras arquitecturas una estructura modular, en la que cada módulo se destina a una tarea específica: toma de decisiones, memorización, aprendizaje, percepción o actividad motora. Finalmente, la arquitectura Soar opta por representar el conocimiento del sistema de forma simbólica.

En el campo de las arquitecturas híbridas o mixtas, las más prometedoras son aquellas en las que los distintos módulos desarrollan sus funciones de forma integrada en las distintas fases del proceso cognitivo, como por ejemplo la ACT-R (siglas del inglés *Adaptative Control of Thought – Rational*, o «Control adaptativo del pensamiento racional») o la CLARION.

La arquitectura CLARION, cuyo funcionamiento básico se recoge en la figura 4, es un diseño de Ron Sun, del Instituto Politécnico Rensselaer. CLARION destaca por dividir el proceso cognitivo entre el implícito y el explícito, una distinción que se corresponde a grandes rasgos con la que hemos establecido entre el aprendizaje inductivo y el deductivo, respectivamente. En CLARION, las tareas cognitivas

Fig. 4



Estructura básica de la arquitectura cognitiva mixta CLARION. El flujo de entrada y salida de datos es en sentido izquierda-derecha.

implícitas son responsabilidad de redes neuronales. El aprendizaje se realiza principalmente «de abajo arriba», es decir, con el ámbito implícito alimentando al explícito, el cual, a su vez, provee de datos etiquetados a los procesos inductivos ulteriores. El ámbito explícito no incorpora de forma acrítica el conocimiento procedente de la inducción, sino que lo considera una hipótesis sujeta a corroboración para la cual echa mano del conocimiento explícito acumulado.

CLARION, al igual que ACT-R, emplea una división entre la memoria procesal, donde se recoge la información fruto de la acción, y la memoria declarativa, que incorpora el conocimiento ya validado. CLARION llama a una y otra *subsistema centrado en la acción* y *subsistema no centrado en la acción,* respectivamente. El papel del primer subsistema es controlar tanto las acciones externas como las internas. La capa implícita de este subsistema está formada por redes neuronales denominadas *redes neuronales de acción,* mientras que la capa explícita se compone de reglas que guían la acción. El papel del *subsistema no centrado en la acción* es el de mantener el conocimiento general. La capa implícita está compuesta por redes neuronales asociativas, es decir, las que generan conocimiento de hechos complejos mediante la asociación de hechos básicos, mientras que la capa inferior está compuesta por las reglas que rigen dicha asociación. El conocimiento se divide, además, en semántico y episódico, donde el conocimiento semántico es generalizado, y el conocimiento episódico es aplicable a situaciones más específicas. A ambos subsistemas se añade un tercero, de *motivación,* encargado de establecer los objetivos que guían las acciones, y el de la *metacognición,* que recoge el elemento reflexivo y que incide, sobre todo, en la evaluación continua de los objetivos a la luz del conocimiento adquirido. Por su coherencia con las teorías más actuales de la psicología cognitiva, CLARION suele emplearse para someter a validación hipótesis psicológicas, e incluso para simular la creatividad.

Además de banco de pruebas, las arquitecturas cognitivas se han usado en la simulación de agentes inteligentes en ámbitos virtuales como juegos de ordenador o simuladores. Y solo cabe esperar que su sofisticación y fidelidad aumenten conforme lo haga la ciencia cognitiva. Con las arquitecturas cognitivas más destacadas, estamos vislumbrando ya, probablemente, la primera mente artificial.

05

EL FUTURO ES HOY. LA IA EN EL MUNDO REAL

Hay tres ámbitos en los que el papel transformador de la IA se muestra con especial contundencia: la automatización de tareas, la predicción del comportamiento y el internet de las cosas.

Cuanto más digitalizada está una actividad, mayor valor le aporta la IA. Por ello, el auge de esta última ha venido de la mano de la madurez de la economía digital experimentada en las dos últimas décadas. No hay ámbito donde su presencia no se haga sentir cada vez más, ya sea en el transporte y la logística, o bien en la salud, el hogar o el comercio. Estamos hablando siempre de IA débil, es decir, de inteligencias circunscritas a dominios concretos. Ahora bien, en la medida en que estas inteligencias de vuelo corto se integren y colaboren de modo creciente, como en el caso del *Internet of Things* (IoT), el resultado puede acabar siendo mucho más que la suma de las partes.

Esta ubicuidad creciente de la IA hace inviable un repaso mínimamente completo de sus aplicaciones presentes o próximas en el corto espacio de estas páginas. En su lugar, vamos a exponer tres estudios de caso de especial interés no solo por la tecnología implicada sino porque nos permiten explorar las consecuencias morales y sociales de la IA. Unas consecuencias que, como no podía ser de otra manera teniendo en cuenta el potencial disruptivo de esta tecnología, son muy importantes.

LA SUSTITUCIÓN DEL SER HUMANO POR MÁQUINAS

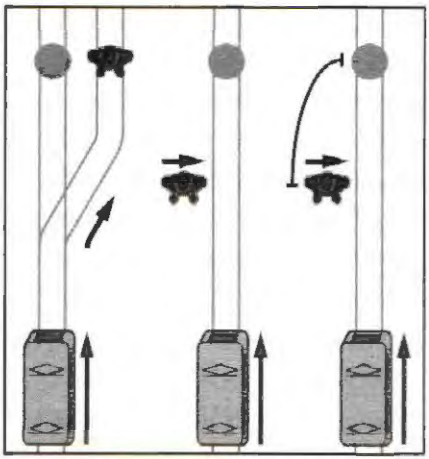
Una de las consecuencias de la digitalización económica es la sustitución del ser humano por máquinas en un número creciente de tareas. Un caso paradigmático de automatización de un proceso considerado extremadamente complejo es la conducción. Un vehículo totalmente autónomo que circule por una carretera estándar con la misma o mayor seguridad que la que ofrece un conductor humano está ya al alcance de la tecnología vigente. Cabe recordar que ya en 2011, el estado norteamericano de Nevada se emplazó a regular la circulación de coches sin conductor.

La ventaja obvia de esta tecnología con respecto a la tradicional es la mayor seguridad vial. Los beneficios del coche autónomo en este ámbito se pueden traducir en un ahorro de hasta 1 240 000 vidas a escala global. Otro claro beneficio sería la posibilidad de optimizar las rutas por vía tecnológica, lo que haría disminuir la emisión de gases contaminantes por una doble vía: la racionalización del tráfico y la reducción del parque total de vehículos.

El papel de la IA en esta revolución es doble. Por un lado, las tecnologías de aprendizaje automático analizan los datos procedentes de los sensores e identifican los objetos y monitorizan su desplazamiento. Por el otro, técnicas de razonamiento automático permiten al ordenador de a bordo planificar los movimientos del vehículo y establecer las mejores rutas.

Ahora bien, este escenario de autonomía total del vehículo plantea importantes dilemas éticos. Imaginemos que la IA del coche se encuentra en la tesitura de frenar bruscamente para evitar atropellar a cinco peatones pero que, si lo hace, volcará y matará al pasajero que transporta. Desde un punto de vista puramente utilitario, la búsqueda del mal menor aconsejaría sacrificar al conductor, pero ¿realmente nos subiremos a un vehículo programado para sacrificarnos, aunque sea en casos extremos? ¿Y quién decidirá previamente qué programación «moral» es la adecuada? ¿Dejaremos que

> EL DILEMA DEL TRANVÍA



— Tres variantes del dilema del tranvía. 1º: Dilema clásico. 2º: El hombre del tejado. 3º El villano.

Hay un célebre experimento mental que pretende clarificar cuestiones morales como la que se planteaba al conductor del vehículo autónomo: el dilema del tranvía. La versión más conocida propone lo siguiente: un tranvía sin control está a punto de atropellar a cinco personas que están atadas a la vía. Hay un botón que acciona un cambio de vía, pero por este camino secundario hay, también, una persona que será atropellada. ¿Debería de pulsarse el botón? La respuesta dependerá de qué enfoque moral es el que se privilegie. El enfoque utilitarista defiende, de forma simplificada, que la decisión correcta es la que conlleva una ganancia mayor de utilidad (entendida como bienestar, felicidad o similar). En cambio, otros enfoques defienden que hay acciones que resultan inmorales sin importar cuánta utilidad generan, como por ejemplo el asesinato. El enfoque utilitarista defendería pulsar el botón; el segundo, lo contrario. La mayoría de las personas maneja principios morales que combinan uno y otro enfoque. En una variante del dilema, la de «el hombre del tejado», el dilema está entre cambiar de vía o empujar a un hombre desde un tejado de forma que al caer sobre la vía detenga el tranvía. A pesar de que el resultado entre ambos dilemas es el mismo (muere una persona) y, por tanto, el cálculo utilitario es el mismo, la gente suele encontrar más problemática la segunda opción pues implica la muerte de alguien no directamente implicado en la situación. Y un dilema de propina: ¿cambiaría nuestra decisión si supiéramos que el hombre del tejado es un criminal?

sea la IA quien decida o querremos que la decisión resida siempre en un ser humano, es decir, en el programador?

Situaciones como esta hacen evidentes que, si se quieren aprovechar todas las potencialidades de la IA, aunque sea la débil, más pronto que tarde habrá que abordar la cuestión de la moralidad de su comportamiento. Para ello, se dispone de dos enfoques posibles. Como veremos, se repite aquí la oposición entre determinismo y adaptación, que, bajo distintas formas (simbolismo contra conexionismo, deducción contra inducción), recorre la historia de la IA. El primer enfoque es del tipo *top-down*, y consiste en determinar por entero las reglas morales de comportamiento del sistema de modo que este se limite a obedecerlas. Entre las dificultades obvias de este enfoque está el diseño de un sistema de reglas que cubra la enorme cantidad de situaciones «morales» con las que nos debemos enfrentar a diario, y que, además, lo haga sin ambigüedades ni zonas grises. Son las dificultades típicas con las que se encuentran los sistemas deterministas tradicionales. A los partidarios de este enfoque les queda la esperanza, por frágil que sea, de que el sistema moral humano, con toda su complejidad, pueda en realidad reducirse a un algoritmo más o menos sencillo todavía por descubrir. El más conocido de los intentos por llevar a cabo una reducción semejante, aunque no se trate más que de un recurso novelístico, son las tres leyes de la robótica propuestas por el escritor Isaac Asimov en su relato *Runaround* (1949):

1. Un robot no hará daño a un ser humano o, por inacción, permitir que un ser humano sufra daño.
2. Un robot debe hacer realizar las órdenes dadas por los seres humanos, excepto si estas órdenes entrasen en conflicto con la 1ª Ley.
3. Un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la 1ª o la 2ª Ley.

En el enfoque *bottom-up*, por el contrario, se programa la IA con unas pocas reglas rudimentarias y se le permite interactuar con el mundo. Utilizando técnicas de aprendizaje automático, desarrollará su propio código moral, adaptándolo a nuevas situaciones a medida que aparezcan. En teoría, a cuantos más dilemas haga frente el sistema, más exhaustivos serán sus juicios morales. Dicho enfoque no está libre de dificultades; por ejemplo, todavía no hemos conseguido diseñar algoritmos de aprendizaje lo bastante complejos como para aprender moralidad.

La amenaza a los puestos de trabajo

La implantación de sistemas de IA débil que sustituyan a los seres humanos en diversas tareas reducirá, como es lógico, la demanda de mano de obra para esas tareas. Los defensores de la implantación sostienen que los empleos abocados a la desaparición son, precisamente, las que no exigían a quienes los desempeñan más que una aplicación rutinaria de una serie de reglas. Al fin y al cabo, poca gente querría trabajar en las cadenas de montaje de principios del siglo pasado, antes de la incorporación masiva de los robots a las fábricas. ¿No se trata aquí de lo mismo, pero trasladado a labores burocráticas? Es tentador responder con otra pregunta: ¿Qué parecido guarda un típico brazo robótico, capaz de una única tarea, con Baxter, el robot de Rodney Brooks que suma a la tradicional fuerza y precisión de los robots la capacidad de aprender una serie de movimientos solo con verlos, así como de transmitir un rango amplio de emociones? Lo cierto es que la versatilidad y potencia de la IA débil, de las que hemos visto numerosos ejemplos, permite a las máquinas suplir a los seres humanos en actividades que son, o al menos eso se creía, mucho más que la aplicación repetitiva de reglas. Y es que la idea misma de establecer una clara separación entre labores susceptibles de ser automatizadas y otras que no es

una tarea fútil en la medida en que la tecnología demuestra ser capaz de automatizar cada vez más y más procesos. En las páginas anteriores hemos examinado en detalle hasta tres de ellas: la traducción, el diagnóstico y, ahora, la conducción. No resulta difícil imaginar otras, como por ejemplo la redacción periodística. Y tengamos por seguro que, si una tarea es automatizable, se automatizará: la ganancia de productividad es demasiado tentadora. ¿Cuándo se detendrá, dicen los más pesimistas, esta espiral destructora de empleo?

Una primera respuesta que darles consiste en recordar que la IA no solo destruye trabajos, sino que también los crea. Por ejemplo, los dedicados al diseño, mantenimiento y gestión de las propias IA; los relacionados con los ordenadores que las tienen que ejecutar y con las máquinas a través de las cuales interaccionan con el mundo y los seres humanos; los de quienes tienen que formar a los expertos en todas esas áreas, y un largo etcétera. A principios del siglo XXI no había consenso acerca del saldo, positivo o negativo, de esta alternancia de creación y destrucción de empleo. En el caso de que acabe siendo lo segundo, cabe la opción de repartir los empleos supervivientes, ya sea trabajando menos horas o bien haciendo que gran parte de la sociedad prescinda del trabajo. Y ese tiempo libre podría dedicarse entonces a la realización personal. Todo sería posible porque la automatización permitiría abaratar enormemente los bienes y servicios. Para abordar escenarios de este tipo, sin embargo, es posible que sean necesarias intervenciones políticas tales como garantizar una renta universal básica o, cuando menos, un amplio reparto de la riqueza generada. El más reputado experto en las consecuencias sociales de las revoluciones tecnológicas, Jeremy Rifkin, escribió al respecto de la automatización:

> A mediados de siglo, nuestros nietos mirarán hacia atrás a la era del empleo masivo con la misma total incredulidad con la que nosotros contemplamos la esclavitud y la servidumbre de otros tiempos. La

> idea misma de que el valor de un ser humano se mide casi exclusivamente por su producción de bienes y servicios y su riqueza material les parecerá primitiva, incluso bárbara, y será considerada como una terrible pérdida de valor humano para nuestra progenie, que vivirá en un mundo altamente automatizado en el que gran parte de lo que se necesita para vivir se produce a bajo coste, o incluso cero, de forma privada y, después, se comparte.

En el futuro vislumbrado por Rifkin, la automatización impulsada por IA trasciende no solo la dimensión tecnológica sino también la puramente económica. Si la explosión de productividad derivada de la Revolución Industrial arrancó a buena parte de la humanidad de la pobreza, la que seguiría de la automatización, comenta Rifkin, acabaría con la idea misma de necesidad y nos conduciría a la abundancia.

La de Rifkin es una visión sin duda optimista del futuro. No es necesario compartirla al cien por cien para estar, no obstante, esperanzado en las posibilidades que la IA nos brinda de transformar el modo en el que trabajamos y producimos.

LA PREDICCIÓN DEL COMPORTAMIENTO

Otro interesante estudio de caso del impacto de la IA en el mundo actual o en el futuro próximo es la que ofrecen los modelos predictivos del comportamiento. Estos modelos han experimentado un auge extraordinario en la estela de la enorme cantidad de datos que generamos diariamente sobre nuestros hábitos de consumo, desplazamiento y demás al interaccionar con dispositivos y programas. Armados con toda esa información, los proveedores de servicios pueden adaptar su oferta a las particularidades de cada usuario hasta unos extremos de ajuste tales que el consumidor pueda llegar a pensar que el algoritmo de IA responsable le ha leído la mente.

A la hora de hacer recomendaciones a los usuarios, el *marketing* digital dispone de dos técnicas principales: el filtraje basado en el contenido y el filtraje colaborativo. Las recomendaciones basadas en el contenido se basan en la afinidad que hay entre las propiedades de los elementos que recomendar, por ejemplo el género de una película o los actores protagonistas. Si el usuario ha mostrado repetido interés en un actor o actriz, por ejemplo, el sistema le recomendaría otras películas protagonizadas por dichos actores. En el filtraje colaborativo, la recomendación es fruto de la comparación de los gustos del usuario con el del resto de los usuarios. De las preferencias mostradas por unos y por otros se dibujan los perfiles respectivos y, si hay coincidencia, el sistema recomienda al usuario lo que sabe que ha gustado a otros usuarios de un mismo perfil. Para llegar a esos perfiles las decisiones de consumo del conjunto de usuarios se deben tamizar en busca de patrones en sus gustos. Se trata de un sistema más flexible y que arroja mejores resultados que el anterior, y como tal es el empleado por gigantes como Amazon o Netflix. Además, sus resultados son tanto mejores cuantos más usuarios se analizan.

> Los ordenadores están logrando para la inteligencia humana lo que la máquina de vapor hizo por la fuerza muscular.
>
> ERYN BRYNJOLFSSON

El establecimiento de patrones en los gustos de los usuarios es una operación análoga al reconocimiento de imágenes que hemos visto al hablar de las redes neuronales. Y al igual que aquellas, los sistemas de recomendación basados en la colaboración necesitan de un conjunto mínimo de datos «etiquetados» por parte del usuario, para determinar su perfil, y de un conjunto mucho mayor de datos también etiquetados por otros usuarios para inducir patrones. Es por ello por lo que solo están al alcance de aquellos servicios que cuentan con un gran número de usuarios. Los algoritmos de estos sistemas son tan sofisticados que son capaces de predecir qué producto le va a gustar a un usuario aunque este no lo haya consumido

nunca. Retomemos el ejemplo del sistema de recomendaciones de películas. El sistema tendría acceso a millones de evaluaciones de películas hechas por los usuarios de la plataforma. En el mundo *offline* la gente se recomienda películas basadas en las opiniones de sus amigos o de personas con perfiles similares. Un sistema de recomendación identifica qué usuarios tienen opiniones semejantes sobre las películas que han evaluado. Cuando queremos predecir si una película le va a gustar a un usuario que aún no la ha visto, analizaríamos si a la gente que tiene evaluaciones similares a este usuario sobre otras películas le ha gustado o no esta película que queremos recomendar.

Extraer la evaluación de los productos o servicios en sistemas de recomendación no es tan sencillo como pueda parecer. La tendencia es que las plataformas no preguntan de forma explícita a los usuarios evaluaciones porque esto requiere de un esfuerzo que el usuario raramente lleva a cabo, con lo que se pierde mucho dato potencial. Además, en muchos ámbitos hay un contraste evidente entre la opinión digamos «pública» y la que indica la actividad real del usuario. Se trata de un contraste similar al que se da entre las encuestas previas a unas elecciones y el resultado final debido a aquellos encuestados que no revelan el sentido real del voto cuando este es para una opción mal vista por la mayoría. Por este motivo, las opiniones de los usuarios son calculadas a partir de parámetros derivados de la actividad implícita del usuario, y no de sus opiniones explícitas. Por ejemplo, si el usuario ha visto una película entera nos indica que le ha gustado mucho más que otro que la ha empezado a ver pero no la ha terminado. O en música, se pueden considerar el número de veces que el usuario ha escuchado una canción, si la canción ha sido incluida en una de sus listas de canciones, etc. En muchas aplicaciones, el *feedback* implícito del usuario nos indica sus opiniones de una forma más real que las opiniones explícitas que el usuario pueda indicar en el sistema.

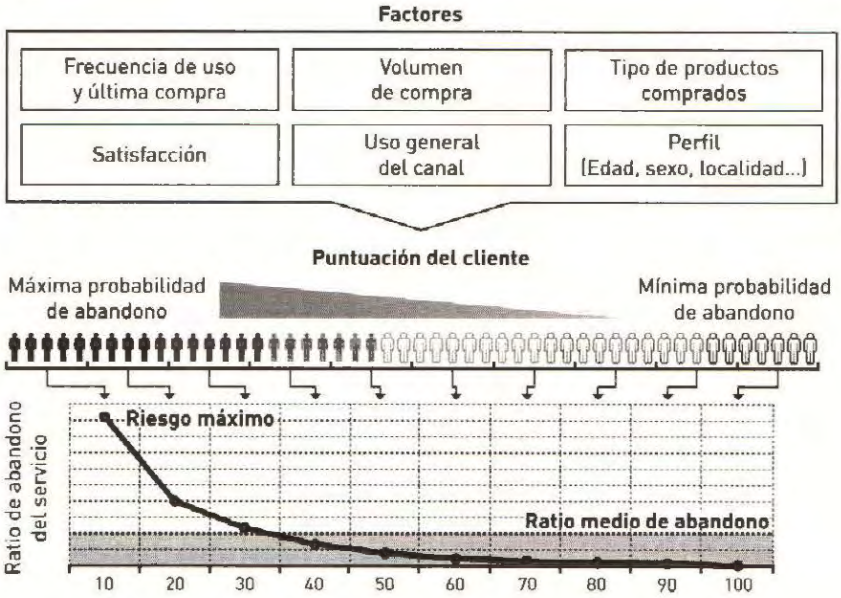
La discrepancia entre el comportamiento implícito y el explícito no siempre obedece a cuestiones conscientes, como en el caso del encuestado que acabamos de ver, sino que en ocasiones responde a impulsos inconscientes que el usuario sería incapaz de explicitar. O también puede deberse a la suma de muchos pequeños comportamientos que, sin ser inconscientes, sí configuran un patrón que el propio usuario podría no ser consciente de que existe. Los modelos predictivos, al ser capaces de detectar pautas en los comportamientos implícitos, son inmunes al autoengaño o a la ignorancia. Modelos de este tipo se emplean habitualmente para prever, por ejemplo, cuándo un usuario está a punto de abandonar un servicio (fig. 1). Armado de esa información, el sistema puede hacerle llegar, de forma automática, una oferta personalizada que le disuada.

La generalización de modelos predictivos capaces de adelantarse a nuestras propias decisiones podría conducirnos, según algunos autores, a un futuro en el que deleguemos de forma voluntaria esas decisiones, incluso las más personales, en algoritmos de IA. El ensayista Yuval Noah Harari, uno de los primeros en abordar esta cuestión, dijo al respecto: «Si un algoritmo te monitoriza todo el tiempo, te conoce mejor que tú. Y puede hacerte recomendaciones sobre con quién tener una cita o casarte, o a quién debes votar». Es importante recalcar que este traspaso de la autoridad desde uno mismo al algoritmo será, en todo caso, voluntario. La cita anterior concluye con un pragmático «al final, será una cuestión empírica: si el sistema funciona bien y te hace buenas recomendaciones, lo escucharás cada vez más.»

TODO CONECTADO CON TODO: EL INTERNET COGNITIVO Y UBICUO

El último caso que vamos a estudiar es el del «internet de las cosas» o IoT, del inglés *Internet of Things*. El concepto fue propuesto por

Fig. 1



Un modelo predictivo del abandono potencial de servicio por parte de un cliente. Se parte de una selección de factores explicativos, por ejemplo la frecuencia de uso del servicio, el volumen de gasto, la satisfacción general del usuario, etc. A partir de la puntuación en cada uno de los factores se elabora una puntuación global del cliente y se lo asigna a un percentil u otro de riesgo de abandono. Los clientes de mayor riesgo en un momento dado son objeto de acciones de fidelización.

Kevin Ashton, del MIT, en 1999, y se refiere a la red global formada por las interconexiones entre dispositivos físicos tales como vehículos conectados, casas inteligentes y, en general, todo aquel capaz de transmitir datos. De su magnitud nos puede dar una idea cálculos que estiman en 30 000 millones el número de dispositivos integrados en el IoT a principios de la segunda década del presente siglo.

La dirección de cada dispositivo que se conecta a internet se gestiona a través del protocolo IPv4, que puede proporcionar direcciones (IP) a aproximadamente 4 000 millones de conexiones.

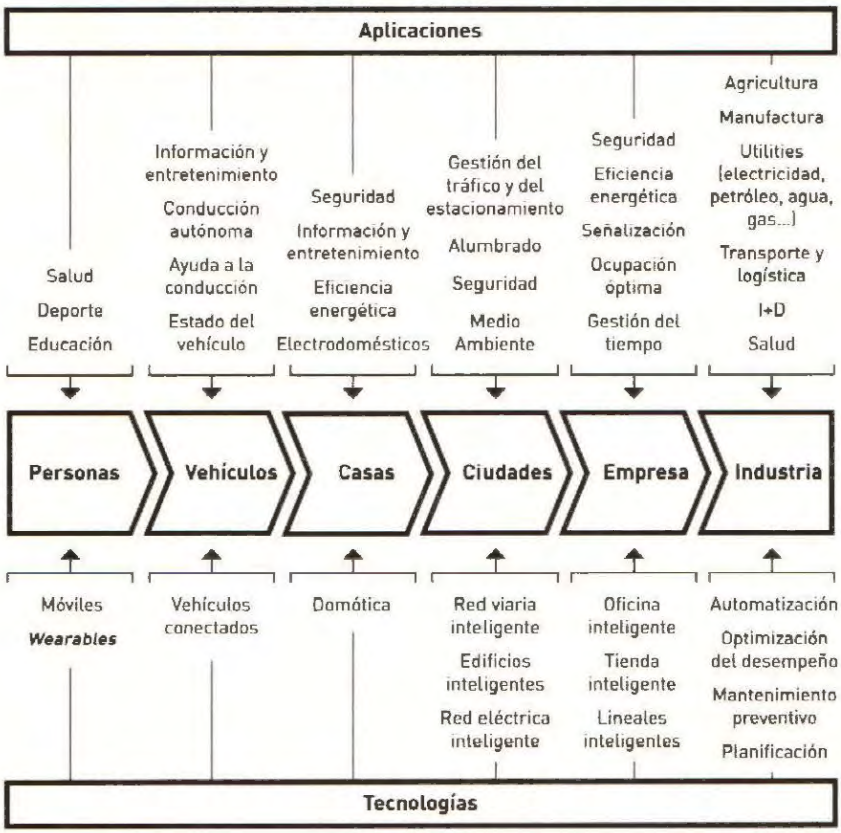
La cifra, debido a la creciente conexión a Internet de ordenadores y *smartphones*, está quedándose corta, y por ello se está diseñando una nueva generación de protocolos, la IPv6. El directorio resultante tendrá suficiente espacio para 3,4 x 1038 (340 undecillones o billones de billones de billones) de direcciones únicas, o lo que es lo mismo: cada átomo que conforma el planeta Tierra podría recibir una dirección única y, a pesar de ello, todavía nos quedarían direcciones para servir a otros cien planetas similares.

> Con el IoT tiene lugar un fenómeno de convergencia en el que bits del ámbito digital se están fusionando con átomos del mundo físico.
>
> Joi Ito, director del Media Lab del MIT

Por el momento, ya disponen de dirección única y se conectan inalámbricamente a internet ordenadores, teléfonos y tabletas, pero el descenso del coste de los sensores está propiciando que también empiecen a hacerlo relojes y pulseras, vehículos, electrodomésticos, y un largo etcétera de objetos que comparten entre sí los datos más diversos: ubicación, velocidad, temperatura, sonido ambiental, fuerza, carga, par de torsión, presión y otras interacciones.

En el futuro que se vislumbra gracias al IoT, en el que todo está conectado con todo, la interacción entre máquinas y seres humanos será desplazada por la que se produzca entre máquinas sin mediación nuestra, la denominada comunicación M2M (siglas del inglés *machine to machine*, «de máquina a máquina»). Las comunicaciones M2M tienen el potencial de tornar más eficientes los procesos existentes en ámbitos tan diversos como la gestión de las ciudades, el comercio o la mayoría de las industrias, desde la sanitaria a la manufactura pasando por la agricultura y las *utilities*. Todo ello sin olvidar las ventajas que ofrece a la escala más reducida de la vida diaria y el hogar (fig. 2). No es extraño que el entorno en el que más se ha adelantado a día de hoy en el IoT sea en el doméstico.

Fig. 2



El internet de las cosas está compuesto de una miríada de objetos interconectados, de electrodomésticos a edificios. Esta interconexión tiene múltiples aplicaciones en ámbitos que van del personal al industrial, y en escalas que viajan de lo doméstico a la producción a escala mundial.

Pero la comodidad personal o la mejora de la eficiencia productiva, con toda su importancia, son solo la punta del iceberg de esta tecnología. Porque de los datos procedentes del IoT, de cuya cantidad a largo plazo nos ponen en la pista las mareantes cifras del protocolo IPv6, se extraerán correlaciones, se detectarán patrones

y tendencias y, en consecuencia, se predecirán hechos. Es el paradigma del *big data* llevado a la enésima potencia.

Si pudieran equipararse estas poderosas herramientas a un ser humano, el IoT serían los sentidos y el sistema nervioso, el *big data*, la información que alberga su cerebro, y los algoritmos de IA que se manejarán para interpretar esta última, la cognición. Por ello, la convergencia de estas tres herramientas se conoce como *internet cognitivo y ubicuo*.

Esta nueva plataforma permitirá gestionar cantidades ingentes de información a la vez que toma decisiones en escalas temporales del orden de los milisegundos. Y lo más importante es que esa información ya no solo se proporciona al usuario humano, sino que se comparte entre las distintas máquinas para que estas adquieran autonomía y, a su vez, el conocimiento entre a formar parte del sistema perceptivo de la Red. A principios de la segunda década del presente siglo, más del 61% del tráfico mundial de internet lo generaban ya cosas, en vez de personas. El auge de la economía digital al que nos referíamos al principio del capítulo es el viento que impulsa la tormenta perfecta de IoT, *big data* y sistemas inteligentes que se anuncia en el horizonte.

Y, entonces, la analogía entre el internet cognitivo y ubicuo y un ser humano individual se quedará corta: el IoT serían los sentidos y el sistema nervioso de todos los seres humanos de la Tierra, el *big data*, toda la información generada por todos y por todo, incluida la almacenada en todos los libros y documentos, y la IA, la capacidad de extraer y gestionar todo ese conocimiento. En este escenario, la ultrainteligencia no sería patrimonio de un único ordenador superpoderoso, esa versión tecnológica del dios tradicional, sino que surgiría de la acción colectiva y armoniosa del ser humano y de la máquina.

LECTURAS RECOMENDADAS

Bengio, Y., Courville, A, Goodfellow, I., *Deep Learnig*, MIT Press, 2017.

Bostrom, Nick, *Superinteligencia: Caminos, peligros, estrategias*, Teell Editorial, 2016.

Copeland, Jack, *Inteligencia artificial*, Alianza Ed., 1996.

Cukier, K., Mayer-Schönberger, V., *Big data. La revolución de los datos masivos.* Turner, 2015.

Minsky, Marvin, *La máquina de las emociones*, Debate, 2010.

Nilsson, Nils, J. *Inteligencia artificial: una nueva síntesis*, McGraw-Hill, 2000.

Norvig, Peter & Russell, Stuart, *Inteligencia artificial: Un enfoque moderno*, Ed. Alhambra, varias ediciones, 2004-2014.

ÍNDICE

ACT-R 120, 122 *ver también* arquitectura cognitiva
agente
basado en objetivos 72
racional 71, 73, 119
reactivo 72
aprendizaje
no supervisado 118-119
supervisado 107-108
árbol
de búsquedas 73-77
de decisión 109
arquitectura cognitiva 22, 72, 119, 121
automatización 126, 130, 131

Baxter 99, 100, 101, 129
big data 8, 39, 41, 63, 138
Boole, George 83
Bostrom, Nick 17, 18, 19, 20, 61
Brooks, Rodney 97, 99, 101, 129
búsqueda
heurística o informada 50, 77
por fuerza bruta 76, 77
restringida 78

cerebro, simulación completa del 21, 32-37
CLARION 120, 121, 122 *ver también* arquitectura cognitiva
cloud computing 64
computación cuántica 30, 31, 64
conexionismo (o conexionista) 9, 10, 44, 47, 51, 54, 59, 64, 120, 129

Deep Blue 10, 61, 62
deep learning 62, 114, 115, 116, 117

embodied cognition 100

Genghis 99, 101
grafo explícito *ver* árbol de búsquedas
grafo implícito (o de espacio de estados) 73, 75
gramática 90, 93, 94

habitación china, experimento de la 42
heurística 50, 65

inteligencia artificial
débil 64, 66, 117, 125, 129
fuerte 64, 66
general (humana) 19, 31, 44
invierno de la 10, 56, 63
test de 59, 65
inteligencia, explosión de 15, 17, 19, 36
Internet of Things (IoT) o Internet de las cosas 10, 64, 123, 125, 134, 135, 136, 137, 138

lenguaje natural, procesamiento (o reconocimiento) del 11, 63, 67, 90, 95, 117
Logic theorist 47-50, 51, 56, 66
lógica difusa 82, 84

machine learning (o aprendizaje automático) 106
McCarthy, John 44, 49, 50
mente corporeizada 21, 102
Minsky, Marvin 44, 50, 54, 55, 56, 57, 58, 113
Moore, ley de 27, 29, 30, 31, 75
Moravec, paradoja de 59
motor de inferencia 84, 85, 86, 90, 105

n-gram, modelo 90, 91, 92
neurona artificial 22, 51, 53, 111
Newell, Allen 9, 47, 49, 50, 120
nouvelle AI 99

operador o función lógica 80, 81, 82, 83

percepción 89, 95, 99, 100, 120
perceptrón 49, 52, 53, 54, 55, 59, 112, 113
planificación 17, 78, 79, 137
predictivo, modelo 131, 134, 135

racionalidad 69, 89, 119
recomendación, sistema de 132, 133
red bayesiana 88, 89
red neuronal 10, 22, 23, 26, 51, 54, 59, 60, 98, 113, 115, 116
convolucional 113, 114, 115
entrenamiento de 98, 113, 115

regresión 107, 108, 109, 110, 111
robot 10, 22, 65, 78, 97, 99, 101, 102, 128, 129
Rosenblatt, Franz 10, 47, 49, 52, 53
Searle, John 42, 64, 66, 117
simbólico, enfoque 47, 50, 58, 97, 120
Simon, Herbert A. 9, 24, 27, 47, 49, 50
sistema experto 22, 84, 85,
Soar 120 *ver también* arquitectura cognitiva
subsunción, arquitectura de 99

tranvía, dilema del 127
Turing, Alan 9, 11, 15, 37, 41-67
Turing, test de 9, 42, 43, 62

ultrainteligencia o superinteligencia 8, 11, 16, 17, 18, 19, 20, 32, 34, 37, 138

vehículo autónomo 10, 126, 127

Watson, sistema informático 10, 24, 62, 63, 64

XOR, problema 56, 59